Universidad Zaragoza
1542

Trabajo Fin de Grado

Clasificación Automática de Documentos utilizando Wikinoticias

Text Categorization using Wikinews

Autor/es

Enrique Salvador Téllez Alonso

Director/es

Dr. Jesús Tramullas Saz

Facultad de Filosofía y Letras

2019

TÉLLEZ ALONSO, Enrique Salvador

Clasificación Automática de Documentos utilizando Wikinoticias [Manuscrito] / Enrique Salvador Téllez Alonso ; dir. Jesús Tramullas Saz. – 2019. – 62 p.: il. col ; 34 cm. – Mecanografiado. – Trabajo Fin de Grado de Información y Documentación de la Universidad de Zaragoza, 2019.

1.Programas de ordenadores. 2. Documentos – clasificación 3. Medios de comunicación social – clasificación I. Tramullas Saz, Jesús, dir. II. Tít

004.81:159.953 + 070.431.2

A mi querida madre

A la familia que se escoge

A mi tutor por su infinita paciencia

A los que ya no están, a los que acaban de llegar y a los que faltan por llegar

Resumen

La meta del presente trabajo es la creación de un clasificador de documentos a través de los artículos de Wikinoticias. Para conseguirlo, ha sido necesario el conocimiento previo del ecosistema de Wikinoticias y el aprendizaje sobre técnicas de minería de datos. Asimismo, ha sido necesaria la creación de un corpus textual procedente de Wikinoticias y el conocimiento de software específico para realizar un caso práctico.

Por otro lado se pretende acercar las técnicas Big Data a la disciplina de las Ciencias de la Información

Palabras clave

Wikinoticias. Minería de Datos. Minería de Textos. Clasificación Automática de Documentos. Aprendizaje Máquina

Abstract

The goal of this paper is the creation of a document classifier with the articles of Wikinews. To achieve this, we needed prior knowledge of the Wikinews ecosystem and it also, learn about data mining techniques. Likewise, It has been needed to create a textual corpus from Wikinews and acquire specific software to make a case study.

On the other hand, it is intended to bring Big Data techniques to the discipline of Information Sciences

Key Words

Wikinews. Data Mining. Text Mining. Text Categorization. Machine Learning

ÍNDICE

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE TABLAS

ÍNDICE DE GRÁFICOS

1. Introducción

El crecimiento de la población mundial con acceso a Internet ha provocado que, en base a su actividad en la Web, se fuesen creando cantidades ingentes de datos. A consecuencia de ello, la cantidad de datos generada ha creado la necesidad de gestionar dicha cantidad de datos. La solución a este problema es la gestión de los datos a través de técnicas Big Data. A modo de definición, se entiende por Big Data "el conjunto de estrategias, tecnologías y sistemas para el almacenamiento, procesamiento, análisis y visualización de conjuntos de datos complejos" (Casas, Nin, y Julbe, 2019, p. 34).

Estas técnicas han sido extrapoladas a diversos campos donde la gestión, transformación, explotación y visualización de los datos, puede resultar útil para extraer patrones o correlaciones a partir de los mismos datos. Para hablar de técnicas de Big Data, en concreto, se habla de minería de datos ya que el término anglosajón, Big Data, hace referencia a la cantidad de datos mientras que la minería de datos es la extracción de los patrones a través de la cantidad de esos datos utilizando herramientas capacitadas para realizar dicha acción.

Existen campos de estudio como la astronomía, la geología, la medicina o el sector financiero donde se necesitan esas técnicas de minería de datos. En lo que nos concierne, las ciencias de la información han adquirido también especial relevancia en este campo pues, como gestores de la información, parece indispensable conocer y utilizar dichas técnicas. Una de las labores frecuentes de un gestor documental es la clasificación de documentos pues, con ayuda de la minería de datos, es posible agilizar el proceso aplicando técnicas que clasifiquen los documentos de forma automática. A modo de definición de lo que es la Clasificación Automática de Documentos, Sánchez-Jiménez (2007, p. 25) ha elaborado la siguiente:

> "Se puede definir la Clasificación Automática de Documentos, también denominada categorización de textos o topic spotting, como la tarea de asignar automáticamente un conjunto de documentos a una o más categorías preexistentes a través de un conjunto de documentos clasificados por expertos sobre los que el sistema lleva a cabo un proceso de aprendizaje supervisado"

1.1. Justificación del trabajo

El conocimiento de las tecnologías de la información y la comunicación (TIC) a lo largo del grado junto con las líneas de investigación ofrecidas para realizar el trabajo fin de grado, han posibilitado que ideara y desarrollara un trabajo que relacionase dos campos de estudio: la Clasificación Automática de Documentos, campo perteneciente a la minería de datos y Wikinoticias, plataforma perteneciente al conglomerado de proyectos Wikimedia.

Wikinoticias centra sus bases en ofrecer una alternativa a los canales de prensa online convencionales, ofertando calidad a su vez a través de revisiones de los artículos que se publican. Esta característica se encuentra acompañada por la condición de que

estos artículos no tienen firma, es decir, cualquier persona que tenga interés puede publicar noticias en esta plataforma. Por ambas características, este portal ha sido seleccionado como fuente principal de datos para experimentar con los mismos.

La necesidad de este trabajo reside en el uso de parte del corpus textual que ofrece Wikinoticias para la elaboración de un clasificador automático de documentos funcional mediante técnicas de minería de datos. Por tanto, con la elaboración de este trabajo se ofrece:

- Un corpus textual para ejercicios futuros.
- Un Clasificador Automático de Documentos basado en técnicas de minería de datos.
- Un trabajo que evalúa la capacidad de Wikinoticias como fuente de datos para la elaboración de trabajos relacionados con la Clasificación Automática de Documentos

1.2. Estado de la cuestión

El concepto de minería de datos, también denominado data mining en su acepción anglosajona, no es reciente. El establecimiento de las bases de datos y las colecciones de datos surgen en los años 60, lo que supone el punto de partida que permite la aproximación al concepto. Más tarde, en los años 70, se crean los primeros sistemas gestores de bases de datos (SGBD) y, durante esa década, se desarrollan campos afines como las bases de datos relacionales o los lenguajes de programación como puede ser Structured Query Language (SQL). Sin embargo, no es hasta finales de los años 80 y principios de los 90 donde es posible hablar de minería de datos como tal, destacando el almacenamiento masivo de datos (data warehousing), facilitando así el análisis de los mismos. Entrados los años 90, el desarrollo de técnicas como el clustering o la clasificación, posibilitaron el análisis de los datos. Esto vendrá reflejado en años posteriores hasta el día de hoy con el concepto de Knowledge Discovery in Databases o KDD (Han, Kamber y Pei, 2012, p. 3)

De forma paralela, la accesibilidad a Internet y su continuo desarrollo permitirían la creación de portales Web cada vez más sofisticados. La actividad en la Web ha dado lugar al incremento del volumen de Internet, provocando, a su vez, la dificultad de la gestión de los datos generados por los internautas. Por contrapartida, este crecimiento también ha posibilitado la creación de una red de comunicación vía Internet, viabilizando el desarrollo de comunidades como Wikimedia, capaces de desarrollar proyectos a gran escala. Como paradigma existe la Wikipedia, siendo esta objeto de análisis en diversos campos por su condición open source. Uno de los usos que se le ha dado a Wikipedia ha sido el aprovechamiento de su extenso corpus textual para realizar labores de Clasificación Automática de Documentos, entre otros varios. (Overell, Sigurbjörnsson y van Zwol, 2007, p. 65)

1.3. Objetivos

Con la elaboración de este trabajo se pretenden conseguir los siguientes objetivos:

Objetivo principal: desarrollar un clasificador automático de documentos funcional con la ayuda de parte del corpus textual de Wikinoticias.

Objetivos específicos:

- Crear un corpus de 800 documentos procedentes de Wikinoticias para experimentar con un clasificador de documentos.
- Crear un corpus de 80 documentos procedentes de Wikinoticias para probar el clasificador creado.
- Analizar y seleccionar el software que mejor se adapte a resolver el problema de la Clasificación Automática de Documentos.
- Analizar y explorar las posibilidades que ofrece Wikinoticias como plataforma de noticias.
- Conocer y aplicar diversas técnicas de Minería de datos.

1.4. Metodología

El trabajo planteado se basa en la realización de experimento en entorno de laboratorio informático. Por lo tanto, se ha seguido un método experimental, en el que se ha establecido una hipótesis de trabajo, y se procedido a comprobar la validez de la misma mediante el desarrollo de un caso.

El método se apoya en tres pilares principales:

A. Estudio sobre Wikimedia

En primer lugar, se realizará un resumen de las distintas plataformas que afectan al ecosistema de Wikimedia. En segundo lugar, el análisis se centrará en Wikinoticias, que será la fuente de datos principal para elaborar el trabajo. Dentro de este análisis se destacará su contenido a través del portal de estadísticas Wikimedia Statistics. Mediante el aporte de diversos gráficos, se procederá a justificar su uso en el trabajo.

B. Estudio sobre minería de datos

En este apartado del trabajo se destacará la minería de datos como ciencia auxiliar. Se repasarán las aplicaciones de la minería en distintos campos, se estudiarán los procesos que integra la minería de datos, se dará una visión global de los algoritmos y técnicas que se aplican y, finalmente, se estudiarán las métricas para evaluar dichos algoritmos y técnicas.

C. Elaboración de un caso práctico de minería de datos

Con el análisis de Wikinoticias como fuente de datos y el repaso del campo de la minería de datos, en esta sección se procederá a realizar un caso aplicado que trate sobre Clasificación Automática de Documentos. Para ello, se extraerán diversos documentos de Wikinoticias para ser probados en la herramienta de minería de datos más idónea.

2. Los wikis semánticos

Para dar sentido a una de las principales secciones del trabajo, es necesario definir lo que son los wikis y, en concreto, lo que son los wikis semánticos. En su forma etimológica, la palabra Wiki viene del hawaiano, que significa rápido. En un principio, la palabra Wiki en su sentido tecnológico-cibernético fue utilizada por primera vez por Ward Cunningham para bautizar a su web denominada Wiki Wiki Web en el año 1995. Acudiendo a una definición técnica de la palabra Wiki, Tramullas (2009, p. 100) ofrece la siguiente: "una wiki es una herramienta y también un producto informativo que han permitido una evolución en la gestión de la información textual en la Web y que se derivan de la gestión de esqueletos de código fuente para la programación". (Tramullas, 2009, p. 99)

Por otro lado se encuentra la concepción del término "semántico". La comprensión de este término compuesto no sería igual sin la aportación de Tim Berners-Lee con su concepción de lo que sería la Web semántica en un futuro. Para Berners-Lee (2006, p.18), la Web semántica sería lo siguiente:

> "The Semantic Web (SW) in an attempt to extend the potency of the Web with an analogous extension of people´s behaviour. The SW tries to get people to make their data available to others, and to add links to make them accessible by link following. So the vision of the SW is as an extension of the Web principles from documents to data. This extension, if it happens and is accepted, will fulfil more of the Web´s potential, in that it will allow data to be shared effectively by wider communities, and to be processed automatically by tools as well as manually."

La creación de la Web semántica permitió la evolución de Internet de tal forma que los productos informativos desarrollarían diversas características en muy poco tiempo. De esta manera, la integración de la Web semántica con los wikis, dio paso al desarrollo de los wikis semánticos. Esta evolución integró ciertas características a los wikis como la capacidad de colaborar. Esta característica cooperativa permite a los diferentes usuarios comunicarse entre sí dentro de la plataforma. Tanto es así que la característica de cooperación en los wikis es una de las condiciones *sine qua non* para el correcto funcionamiento de los mismos. En relación a esta característica indispensable, Tramullas (2009, p. 103) ofrece la siguiente idea: "los wikis facilitan un modo de trabajo que se fundamenta en la edición y modificación en colaboración de documentos, en un entorno informático distribuido". Como añadido, Pacuit y Parikh (2006, p. 1-2) indican que estas aplicaciones se encuentran dentro de un conjunto de herramientas de software denominadas social software, cuyas características principales son:

- Aumentar la capacidad de las personas para comunicarse.
- Aumentar la capacidad para colaborar.
- Compartir información y conocimientos.
- Crear comunidades virtuales.

Esta concepción de cooperación fue tomando cada vez más fuerza a medida que iba transcurriendo el tiempo. Gracias a la comunicación entre internautas, pronto empezaron a aparecer diversos proyectos entre la comunidad. Como paradigma de wikis colaborativas existe la Wikipedia. Fundada por Jimmy Wales y Larry Sanger en el año 2001, la Wikipedia se basó en la plataforma wiki de Ward Cunningham que permitía depositar documentos en la Red de forma cooperativa (Saorín, 2012, p. 11). A pesar de que Wikipedia figure como piedra angular de los proyectos cooperativos, no es el único ya que la Fundación Wikimedia (WMF, Wikimedia Foundation) ha creado diversos proyectos de diferente temática.

3. La Fundación Wikimedia

La Fundación Wikimedia (2019) se encarga de ofrecer diferentes servicios de información a la comunidad de Internet de forma abierta, colaborativa y gratuita. Para introducir este epígrafe, se hará uso de la definición propia de la Fundación Wikipedia sobre sí misma, la cual recoge lo siguiente:

> "La Fundación Wikimedia, una organización sin fines de lucro, provee la infraestructura esencial para el conocimiento libre. Nosotros alojamos Wikipedia, la enciclopedia libre en línea, creada y editada por voluntarios y voluntarias alrededor del mundo así como otros proyectos comunitarios importantes. Damos la bienvenida a cualquiera que comparta nuestra visión para unirse, recolectar y compartir el conocimiento que represente por completo la diversidad humana."(Fundación Wikimedia, 2019).

Uno de los temas que más interés suscita es su forma de financiación. De forma aclaratoria, la Fundación Wikimedia se financia a base de donaciones públicas con la meta de proporcionar un servicio al alcance de cualquier persona en el mundo. Por otra parte, la Fundación Wikimedia apoya de forma económica a eventos de investigación y a la elaboración de contenidos muy específicos para seguir fomentando la creación de cultura y el acceso a la información (Saorín, 2012).

3.1. Proyectos de la Fundación Wikimedia

La cantidad de proyectos impulsados por la Fundación Wikimedia es considerablemente grande. Tanto es así que autores como Fontanills (2012, p. 26) hablan de wikimediaesfera para concebir al conglomerado de proyectos que la Fundación Wikimedia tiene en su haber. A día de hoy, los proyectos de la Fundación Wikimedia se ven segmentados por cinco principales vertientes:

- Referencia.
- Colecciones.
- Tecnología.
- Guías.
- Colaboración.

Dentro de cada una de estas vertientes se encuentran los proyectos respaldados por esta fundación. A modo ilustrativo, se incluye una serie de tablas que ilustran y describen los mismos:

Tabla 1: proyectos de referencia

Proyecto	Descripción
Wikipedia	Wikipedia es la enciclopedia libre y colaborativa escrita en 300 idiomas por voluntarios alrededor del mundo.
Wikilibros	Wikilibros busca construir una colección de recursos libres en e-book incluyendo libros de texto, versiones comentadas, guías instructivas y manuales.
Wikcionario	Wikcionario es el diccionario multilingüe libre. El proyecto busca describir todas las palabras de todos los idiomas. Incluye recursos del idioma como tesauros, guía de rimas y estadísticas del idioma.
Wikiquote	Wikiquote es una colección online de citas referenciadas de personas destacadas y trabajos creativos en más de 75 idiomas. Incluye proverbios, mnemotecnias y lemas.

Tabla 2: proyectos sobre colecciones

Proyecto	Descripción
Wikimedia Commons	Wikimedia Commons, la biblioteca de ilustraciones, fotografías, dibujos videos y música más grande del mundo.
Wikisource	Wikisource es una biblioteca de licencias libres, textos y documentos históricos incluyendo poesía, documentos de gobierno, constituciones de muchos países y literatura en general.
Wikiversidad	Wikiversidad está dedicada a recursos de aprendizaje, proyectos educativos e investigación para su uso en todos los niveles, tipos y estilos de educación.
Wikiespecies	Wikiespecies es la base de datos de especies por taxonomía que incluye representantes vivos y fósiles de Animalia, Plantae, Fungi, Bacteria, Archaea, Protista y otras formas de vida.

Tabla 3: proyectos sobre tecnología

Proyecto	Descripción
Wikidata	Wikidata actúa como el almacenamiento central de información estructurada para los proyectos Wikimedia. Estructurando datos en un formato legible para máquinas los hace más sencillos de visualizar, buscar, editar, curar, usar y reusar archivos.
MediaWiki	MediaWiki es el wiki software libre y de código abierto que cualquiera puede usar y desarrollar. Es la plataforma en la que están hechos los proyectos Wikimedia.

Tabla 4: proyectos sobre guías

Proyecto	Descripción
Wikiviajes	Wikiviajes tiene como fin crear la guía de viajes más grande, completa y actualizada en todo el mundo.
Wikinoticias	Wikinoticias provee contenido libre y es una alternativa a los sitios comerciales de noticias con artículos verificados y revisados por colegas.

Tabla 5: proyectos colaborativos

Proyecto	Descripción
Meta-Wiki	Meta-Wiki es un proyecto usado como el nodo centrar la organización y coordinación de varias tareas como discusiones que afectan a múltiples wikis o planear los próximos eventos.

Es posible observar que no solamente existe Wikipedia, sino que la Fundación Wikimedia se ha encargado de ofertar servicios gratuitos variados al público de internet con el desarrollo de varios proyectos. De todos los proyectos ilustrados anteriormente, el que interesa para el desarrollo de este trabajo es Wikinoticias. Por esa razón, el siguiente epígrafe vendrá dedicado al mismo.

4. Wikinoticias

En palabras de Saorín (2012, p. 15), Wikinoticias es "una fuente colaborativa de elaboración y difusión de noticias, complementaria para muchos temas que no tienen cabida en la Wikipedia". Por otro lado y de forma más extensa, existe la definición que ofrece Abolhosen (2017, p. 43-44) que afirma lo siguiente:

> "En particular, Wikinews es un experimento temerario, ya que se propone como una especie de agencia noticiosa colaborativa. Aquí, como el material son noticias y la actualidad, hay presiones temporales que van a determinar la validez de la información y la eficiencia del servicio, a diferencia de la Wikipedia donde las correcciones, contribuciones y ediciones a un artículo pueden llevarse a cabo durante años y no hay una fecha límite de entrega. Además, el compromiso fundamental es con la neutralidad, una meta difícil de cumplir cuando los articulistas son desconocidos."

4.1. Contenido de Wikinoticias

El contenido principal de la plataforma son artículos de noticias que se presentan como una alternativa a los canales de noticias habituales o convencionales. Al igual que Wikipedia, Wikinoticias se encuentra disponible en 24 idiomas entre los que destacan: francés, alemán, portugués, español o ruso, entre otros. Dicha característica idiomática repercute en la cantidad de artículos pues cada idioma tiene su propio volumen de artículos. A modo de ejemplo, se ilustrará una tabla con los principales idiomas disponibles en relación con la cantidad de artículos disponibles a fecha de 9 de noviembre de 2019.

Tabla 6: número de artículos por idioma

Idiomas	**Número de artículos**
Español	11.000 + artículos
Inglés	21.000 + artículos
Francés	21.000 + artículos
Alemán	13.000 + artículos
Portugués	13.000 + artículos

Asimismo, su contenido se encuentra soportado por la herramienta de software MediaWiki, siendo esta un wiki engine muy común para mantener wikis (Tramullas, 2009, p. 113)

4.2. Organización del contenido

La organización del contenido es similar al que ofrece Wikipedia puesto que alberga una amplia variedad de categorías y subcategorías. Para simplificar la organización de Wikinoticias, los artículos obedecen a tres grandes distribuciones: región, tema y fecha

de publicación. A modo de ilustración, se procederá a representar estas modalidades mediante la siguiente tabla:

Tabla 7: categorías de Wikinoticias

Artículos por tema	**Artículos por región**	**Artículos por fecha**
Arte, cultura y entretenimiento	África	Por año de publicación
Ciencia y tecnología	América	
Clima	América del Norte	
Deportes	América Central	
Desastres y accidentes	América del Sur	
Ecología	Asia	
Economía y negocios	Europa	
Judicial	Medio-Oriente	
Obituario	Antártica y Oceanía	
Política		
Salud		
Sociedad		

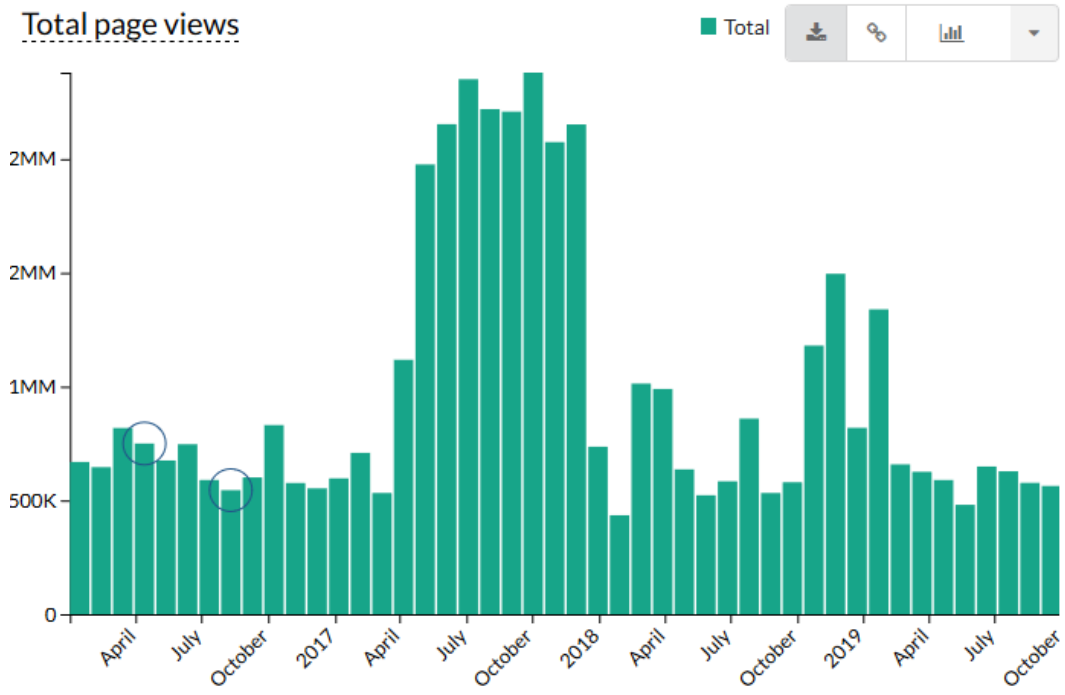
4.3. Wikinoticias como objeto de análisis

Es evidente la carencia de trabajos sobre Wikinoticias pues esta plataforma no ha suscitado suficiente interés entre la comunidad. Este desinterés viene dado por la propia concepción de Wikinoticias sobre los estudiosos en la materia pues varios ponen en duda el futuro del proyecto. A modo de ejemplo, Saorín (2017, p. 195) afirma lo siguiente: "mientras Wikipedia ha demostrado sólidamente su proposición de valor, Wikinews arrastra numerosas dudas sobre su viabilidad". Esta idea no se encuentra aislada, sino que viene fundamentada por trabajos anteriores realizados por otros autores como Keegan (2013) y Bruns (2006) que achacan a la plataforma de estar sufriendo un posible "estancamiento" o "fracaso".

4.3.1. Analizando Wikinoticias con Wikimedia Statistics.

Para contrastar la información proporcionada, es posible recurrir a Wikimedia Statistics. Este portal provee métricas sobre los diversos proyectos de la wikimediasfera, revelando datos sobre lectura, contribución y contenido desde los inicios de cada iniciativa. Algunas de las estadísticas sobre Wikinoticias son las siguientes

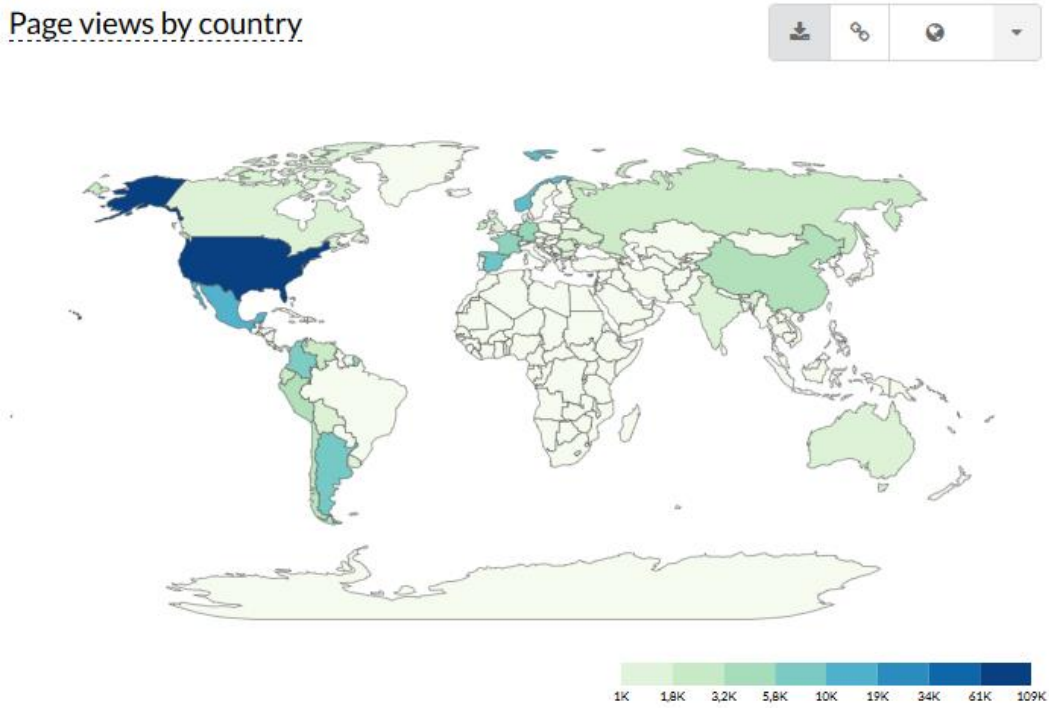
Ilustración 1: número total de visitas en Wikinoticias



Fuente: https://stats.wikimedia.org/v2/#/es.wikinews.org/reading/total-page-views/normal||2016-01-01~2019-11-01|~total|

Esta estadística representa el total de visualizaciones en las páginas de Wikinoticias en los últimos tres años. En este gráfico es posible ver repuntes a lo largo de los tres años pero las visualizaciones poseen cierta estabilidad. El mayor repunte apreciable en el gráfico es entre los meses de abril y diciembre del año 2017, alcanzando cifras superiores a dos millones de visualizaciones en algunos casos.

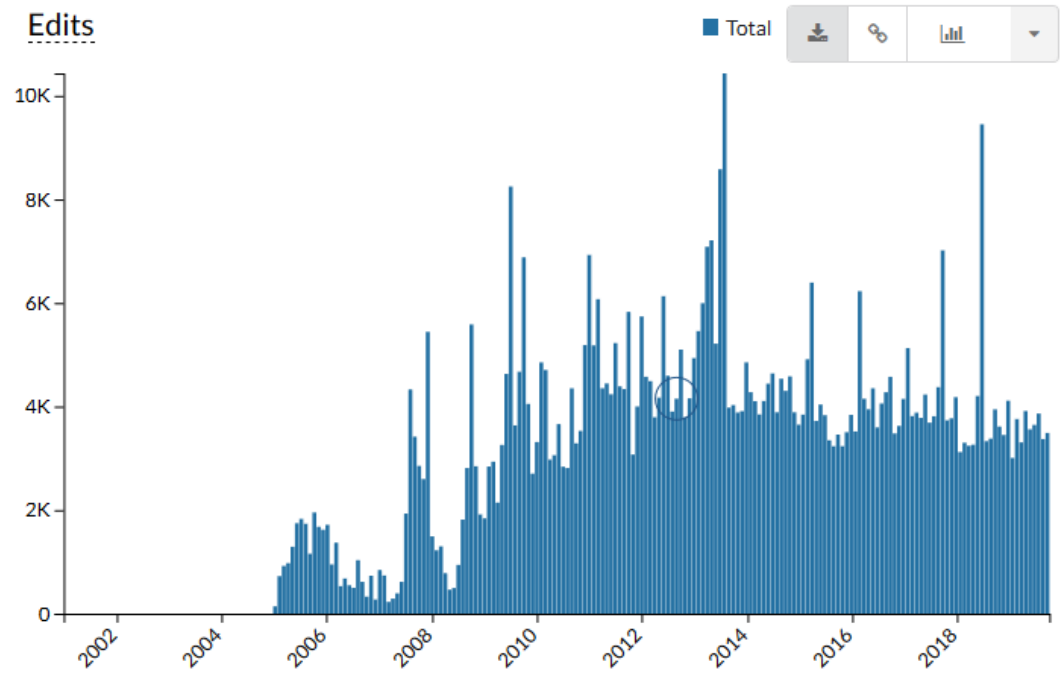
Ilustración 2: número de visitas por país



Fuente: https://stats.wikimedia.org/v2/#/es.wikinews.org/reading/page-views-by-country/normal||2019-10-01~2019-11-01|~to

Este mapa representa el número de visualizaciones por país en el mes actual, en este caso, noviembre de 2019. Siguiendo la leyenda del mapa, es posible observar que el país que más visitas obtiene es Estados Unidos, con 109K (a causa de su comunidad hispanohablante). Le siguen países como México, con 18K, Noruega, con 15K, España, con 12K, Argentina, con 11K y finalmente Colombia con 10K. Los países restantes han tenido una cifra inferior a los 10K con lo que se respecta a Wikinoticias en español.

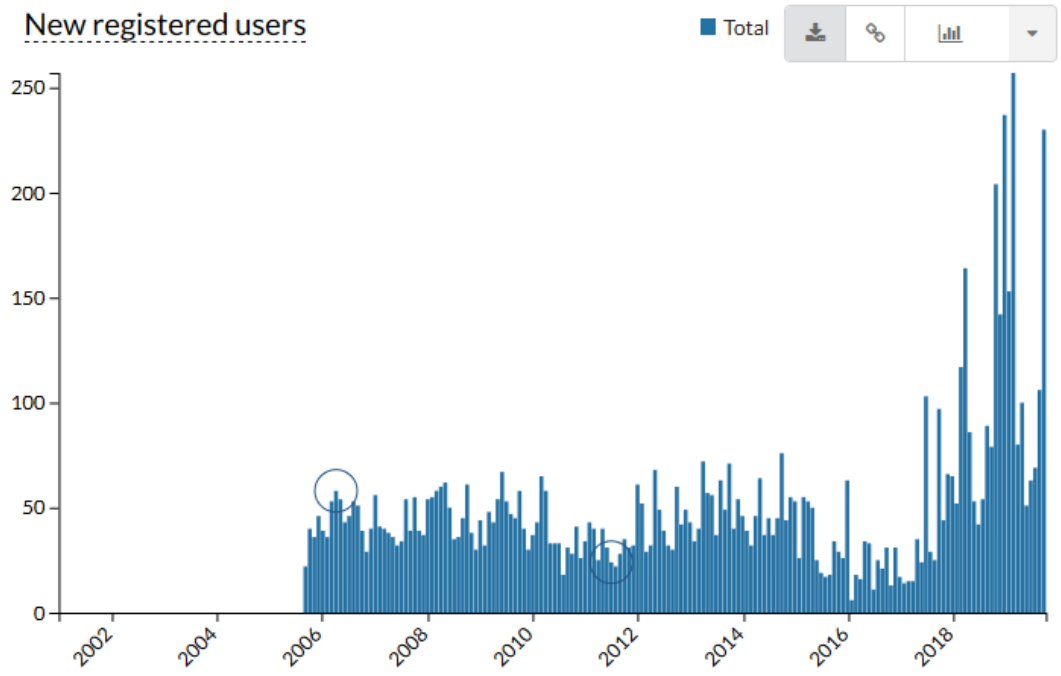
Ilustración 3: número de ediciones



Fuente: https://stats.wikimedia.org/v2/#/es.wikinews.org/contributing/edits/normal||2001-01-01~2019-11-01|~total|

Esta gráfica representa la cantidad de ediciones desde la creación de Wikinoticias. Es observable que el proyecto iniciado en 2005 sufre de ciertos descensos en la edición de sus artículos en años como 2006 y 2007. No es hasta el año 2009 donde es observable un crecimiento en las ediciones y cierta estabilidad en los años venideros, llegando así hasta 2019. Asimismo, dentro de ese período de 10 años, es observable que hay momentos puntuales donde las ediciones pueden llegar a alcanzar cifras cercanas o iguales a 10K. Asimismo, en ese lapso de tiempo es posible observar ciertas depresiones en algunos meses puntuales donde las cifras se encuentran por debajo de 4K. A modo de conclusión, las ediciones se mantienen estables pues a partir del año 2010 las ediciones no descienden de 2K.

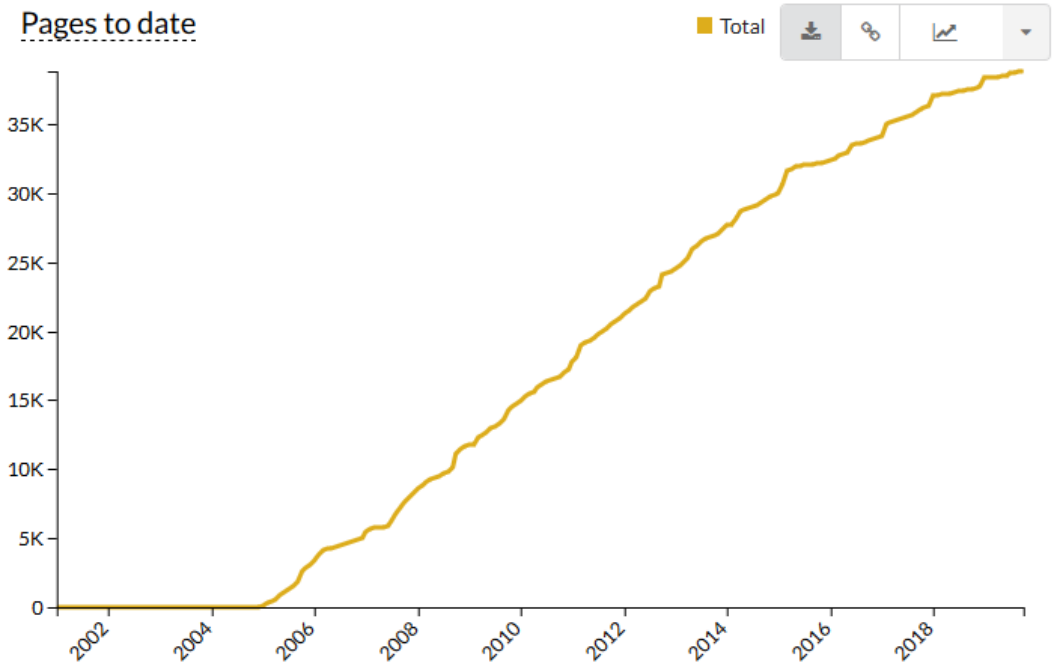
Ilustración 4: nuevos usuarios registrados



Fuente: https://stats.wikimedia.org/v2/#/es.wikinews.org/contributing/new-registered-users/normal||2001-01-01~2019-11-01|~total|

Esta grafica muestra el volumen de nuevos usuarios registrados desde la creación de Wikinoticias. Es posible observar que existe estabilidad desde finales de 2005 hasta finales de 2015, donde sufre un descenso hasta 2017, alcanzando cifras inferiores a 25 usuarios registrados por mes. Sin embargo, a partir del año 2017 es posible observar un crecimiento muy pronunciado, llegando a alcanzar una cifra mayor a 250 usuarios registrados en un mismo mes.

Ilustración 5: número de páginas creadas hasta la fecha



Fuente: https://stats.wikimedia.org/v2/#/es.wikinews.org/content/pages-to-date/normal||2001-01-01~2019-11-01|~total|

Esta grafica muestra la creación de páginas en Wikinoticias desde su inicio. En esta métrica es posible ver un crecimiento exponencial a través de los años pues desde su inicio no ha parado de crecer, llegando a alcanzar una cifra superior a 35K (en concreto, 38.738 páginas).

4.3.2. Evaluación de Wikinoticias.

Acorde con lo que indica Saorín (2012, p. 65-66) en su trabajo "Wikipedia de la A a la W", los proyectos de gestión colectiva, como Wikinoticias, están sujetos a ciertas limitaciones. A pesar de que el autor se centra en Wikipedia, dichas limitaciones pueden afectar a cualquier proyecto de gestión colectiva. Las limitaciones son las siguientes:

1. No alcanzar suficiente masa crítica de participación y contenidos para adquirir relevancia.
2. Rebasar la cantidad de información que puede gestionarse productivamente con la aplicación combinada de las tecnologías y métodos de trabajo utilizados.
3. Dificultad para llegar a acuerdos en temas complejos o sensibles.
4. Pocos usuarios activos.
5. Dificultad de recompensar por el esfuerzo de colaboración.
6. Dificultad de elaborar nueva información que integre el resultado del proceso participativo.
7. Incoherencia con otros sistemas de relación y organización.

En lo que a Wikinoticias respecta, no es comparable a un proyecto como puede ser Wikipedia pero, actualmente, es un proyecto que está operativo, que se mantiene y que todavía continúa creciendo y desarrollándose. Apoyándome en los gráficos presentados anteriormente, es visible que el proyecto tiene algunas limitaciones para crecer y que, inclusive, tiene ligeros descensos en cuanto a desarrollo. Sin embargo, en estos últimos años ha disfrutado de un avance en comparación con años anteriores. Remitiéndome a los datos, a día de hoy, Wikinoticias posee una cantidad superior a los 11.440 artículos en español. Asimismo, la edición de los artículos es constante por lo que es una plataforma que tiene soporte por parte de los editores. Para justificar su uso en este trabajo, se realizara una síntesis de las razones por las cuales se utiliza como fuente de datos:

- Por su condición open source.
- Por la revisión de los artículos que se publican.
- Por su variedad en cuanto a categorías y subcategorías.
- Por su mantenimiento.
- Por el volumen de su corpus textual.

A pesar de la escasez de trabajos sobre Wikinoticias, muchos de los artículos centrados en la Wikipedia como objeto de estudio pueden servir como punto de partida para poder enfocar el trabajo sobre Wikinoticias. A modo de ejemplo, es posible destacar el trabajo de Tramullas en el año 2015 titulado “La Wikipedia como objeto de investigación”. En él se hace una revisión de los diversos usos de la Wikipedia, destacando el corpus textual de la Wikipedia entre otros. Para ser más concisos precisos, son los autores Okoli y Schabram (2009) junto a Taraborelli (2013) los que toman como punto de convergencia el corpus textual de Wikipedia.

Por otro lado y, enfocando el trabajo hacia la importancia del corpus de Wikipedia, es posible destacar un reciente artículo elaborado por Mehdi, Okoli, Mesgari, Nielsen y Lanamäki (2017) titulado “Excavating the mother lode of human-generated text: A systematic review of research that uses the Wikipedia corpus”. En este trabajo, Mehdi et al. (2017) realizan una revisión de los productos informáticos capaces de explotar el corpus de Wikipedia para darle uso en diversos campos de las ciencias de la computación. Es posible destacar campos como:

- Recuperación de la Información (Information Retrieval).
- Procesamiento Natural del Lenguaje (Natural language processing).
- Construcción de ontologías (Ontology building).

Es una realidad que el corpus textual de Wikipedia no es comparable con el de Wikinoticias, sobre todo si se tiene en cuenta el volumen de artículos que esta gran enciclopedia posee. Precisamente, es el volumen del corpus de Wikipedia lo que posibilita la realización de trabajos tan exhaustivos sobre el tema. Estos trabajos que hacen uso del corpus textual de Wikipedia, en muchas ocasiones están centrados en labores de minería de datos.

5. Minería de datos

Para introducir esta disciplina, se mostraran diferentes definiciones de la misma. La primera la ofrecen Han, Mannila y Smith (2001, p. 1) que conciben la minería de datos de la siguiente forma: "Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the date in novel ways that are both understable and useful to the data owner".

Recurriendo a su acepción inglesa, data mining es concebido como el "minado o extracción de datos". Una aproximación más extensa es la que ofrece Tramullas (1997, p. 75) que concibe el significado de esta disciplina y la define de la siguiente forma:

> "Nos encontramos ante una metáfora, en el sentido de considerar a las grandes bases de datos como minas, dentro de las cuales es necesario localizar los filones que contienen los materiales preciosos (de hecho, en data mining se habla a menudo de nuggets, o pepitas de oro u otro metal precioso). Durante esa actividad de mining se localizan las nuggets, pero también los patrones que permiten localizar esas valiosas pepitas."

Desde otro punto de vista, la minería de datos puede ser definida como "un intento de buscarle sentido a la explosión de información que actualmente puede ser almacenada" (Mitra y Acharya, 2003, p. 1). El concepto de la minería de datos es muy amplio, tanto es así que su campo de acción lo es de la misma forma. La versatilidad de esta ciencia de la computación hace que sea muy fácil de implementar en otros campos. Para ilustrar su versatilidad, se mostrarán sus distintas aplicaciones a continuación.

5.1. Aplicaciones de la minería de datos

El avance del propio campo junto con el acercamiento de las tecnologías a otras áreas ha permitido que dicha disciplina sea aplicable en otros terrenos. A modo de ejemplo, Riquelme, Ruiz y Gilbert (2006, p. 14) han elaborado una clasificación de áreas donde es posible aplicar técnicas de minería de datos. La clasificación es la siguiente:

- Comercio y banca: segmentación de clientes, previsión de ventas, análisis de riesgo.
- Medicina y farmacia: diagnóstico de enfermedades y la efectividad de los tratamientos.
- Seguridad y detección de fraude: reconocimiento facial, identificaciones biométricas accesos de redes no permitidas, etc.
- Recuperación de información no numérica: minería de texto, minería web, búsqueda e identificación de video, voz y texto, bases de datos multimedia.
- Astronomía: identificación de nuevas estrellas y galaxias.
- Geología, minería, agricultura y pesca: identificación de áreas de uso para distintos cultivos o pesca de explotación minera en bases de datos de imágenes o satélites.

- Ciencias Ambientales: identificación de modelos de funcionamiento de ecosistemas naturales y/o artificiales para mejorar su observación, gestión y/o control.
- Ciencias Sociales: Estudio de los flujos de la opinión pública. Planificación de ciudades: identificar barrios con conflicto en función de valores sociodemográficos.

5.2. Procesos en minería de datos

En palabras de Tramullas (1997, p. 77) "La aplicación de un proceso de data mining sobre un conjunto de datos es una tarea compleja, que debe estar relacionada con la organización en la que se aplique."Al tratarse de una tarea que alberga cierta complejidad, numerosos autores han desarrollado un proceso específico para extraer la información que se precisa de los datos. A modo de ejemplo Molina (2002, p. 4) ha establecido cuatro fases principales:

1. Determinación de los objetivos: trata de la delimitación de los objetivos que el cliente desea bajo la orientación del especialista en data mining.
2. Preprocesamiento de los datos: se refiere a la selección, la limpieza, el enriquecimiento, la reducción y la transformación de las bases de datos. Esta etapa consume generalmente alrededor del setenta por ciento del tiempo total de un proyecto de data mining.
3. Determinación del modelo: se comienza realizando unos análisis estadísticos de los datos, y después se lleva a cabo una visualización gráfica de los mismos para tener una primera aproximación. Según los objetivos planteados y la tarea que debe llevarse a cabo, pueden utilizarse algoritmos desarrollados en diferentes áreas de la Inteligencia Artificial.
4. Análisis de los resultados: verifica si los resultados obtenidos son coherentes y los coteja con los obtenidos por los análisis estadísticos y de visualización gráfica. El cliente determina si son novedosos y si le aportan un nuevo conocimiento que le permita considerar sus decisiones.

Los puntos 2, 3 y 4 del proceso anterior pertenecen expresamente a la fase de data mining como tal. De forma más desarrollada, Riquelme et al. (2006, p. 13-14) han establecido las tareas más habituales a desarrollar en la fase de data mining. Las tares son las siguientes:

- Clasificación: clasifica un dato dentro de una de las clases categóricas predefinidas.
- Regresión: el propósito de este modelo es hacer corresponder un dato con un valor real de una variable.
- Clustering: se refiere a la agrupación de registros, observaciones, o casos en clases de objetos similares. Un clúster es una colección de registros que son similares entre sí, y distintos a los registros de otro clúster.

- Generación de reglas: aquí se extraen o generan reglas de los datos. Estas reglas hacen referencia al descubrimiento de relaciones de asociación y dependencias funcionales entre los diferentes atributos.
- Resumen o sumarización: estos modelos proporcionan una descripción compacta de un subconjunto de datos.
- Análisis de secuencias: se modelan patrones secuenciales, como análisis de series temporales, secuencias de genes, etc. El objetivo es modelar los estados del proceso, o extraer e informar de la desviación y tendencias en el tiempo.

5.3. Algoritmos y técnicas de minería de datos

Para explicar los algoritmos y técnicas que se utilizan en la minería de datos, es necesario tener en cuenta el concepto de machine learning. El machine learning (ML) o en su denominación en español, aprendizaje máquina, es una variante de la disciplina de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a los ordenadores aprender (Sebastiani, 2002). Para describir esta rama de la inteligencia artificial de una forma más completa, es posible hacer uso de la definición elaborada por Jordan y Mitchell (2015, p. 255) que resume lo siguiente:

> "Machine learning is a discipline focused on two interrelated questions: How can one construct computer systems that automatically improve through experience? and What are the fundamental statistical-computational-information-theoretic laws that govern all learning systems, including computers, humans, and organizations? The study of machine learning is important both for addressing these fundamental scientific and engineering questions and for the highly practical computer software it has produced and fielded across many applications."

Este aprendizaje automático depende de los datos de los cuales se alimenta el software seleccionado. Esto posibilita que se genere un modelo el cual pueda ser capaz interpretar información nueva. Según Murphy (2012), en la disciplina del machine learning se diferencian dos grandes grupos: el aprendizaje supervisado o predictivo y el aprendizaje no supervisado.

5.3.1. Aprendizaje supervisado o predictivo.

El aprendizaje es supervisado o predictivo "cuando la máquina aprende no sólo de los propios datos finales (inputs) sino que es posible darle modelos o datos adicionales ya categorizados (outputs) para que el aprendizaje sea mucho más fiable" (Calderón, Caro y Lorenzo, 2016, p. 628) . De una forma más concisa precisa, para hablar de aprendizaje supervisado se habla de los métodos capaces de realizar dicho aprendizaje. Para representar lo anterior, Gironés, Casas y Minguillón (2017, p. 39) han elaborado una definición de lo que son los métodos supervisados:

> "Los métodos supervisados (supervised methods) son algoritmos que basan su proceso de aprendizaje en un juego de datos de entrenamiento convenientemente etiquetado. Por etiquetado entendemos que para cada ocurrencia del juego de

> datos de entrenamiento conocemos el valor de su atributo objetivo o clase. Esto le permitirá al algoritmo poder deducir una función capaz de predecir el atributo objetivo para un juego de datos nuevo."

5.3.2. Aprendizaje no supervisado.

Para Gironés et al. (2017, p. 39) los métodos no supervisados son aquellos que:

> "Los métodos no supervisados (unsupervised métodos) son algoritmos que basan su proceso de entrenamiento en un juego de datos sin etiquetas o clases previamente definidas. Es decir, a priori no se conoce ningún valor objetivo o de clase, ya sea categórico o numérico."

Calderón et al. (2016, p. 629) proporcionan otra definición que establece que "a diferencia del aprendizaje supervisado, el no supervisado utiliza procedimientos inductivos, extrayendo conocimiento sólo de los datos." Tanto el aprendizaje supervisado como el no supervisado dependen de los algoritmos que deben aplicarse en el caso que se adecuado. El aprendizaje supervisado tendrá unos algoritmos concretos y el no supervisado también tendrá los suyos propios. Para sintetizar dicha información, se presenta una tabla que pueda agrupar los algoritmos propios del aprendizaje supervisado y del no supervisado:

Tabla 8: técnicas de machine learning

Algoritmos supervisados	Algoritmos no supervisados
K vecinos más cercanos	K-means y derivados
Maquinas de soporte vectorial	Agrupamiento jerárquico
Redes neuronales	
Árboles de decisión	
Naïve Bayes (modelo probabilístico)	

Por otro lado, estas técnicas necesitan ser evaluadas una vez son aplicadas. Para ello, se acude a diversas métricas que posibilitan la evaluación de los algoritmos. Por ende, esta evaluación permitirá seleccionar el mejor algoritmo para cada caso en el cual este sea aplicado.

5.4. Métricas de evaluación de algoritmos

Existe una gama muy amplia de métricas que permiten evaluar algoritmos en la minería de datos. Dichas métricas se encuentran estrechamente relacionadas con los problemas propios de la minería de datos (véase el apartado procesos en minería de datos). A modo de ejemplo, Gironés et al. (2017, p. 77) realizan un apunte sobre esto: "Las métricas para realizar este tipo de evaluación dependen, principalmente, del tipo de problema con el que se está lidiando. En este sentido, existen métricas específicas para problemas de clasificación, regresión y agrupamiento." Asimismo, muchas de estas métricas son derivadas de técnicas como las matrices de confusión. Gironés et al. (2017, p. 78) definen la matriz de confusión de la siguiente forma:

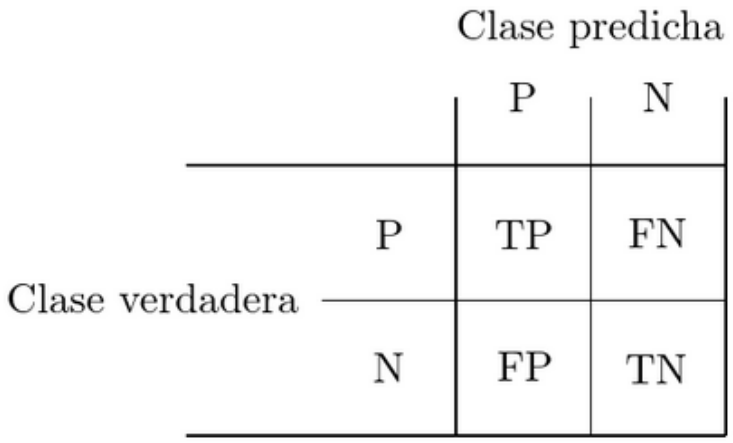
> "La matriz de confusión (confusion matrix) presenta en una tabla una visión gráfica de los errores cometidos por el modelo de clasificación. Se trata de un modelo gráfico para visualizar el nivel de acierto de un modelo de predicción. También es conocido en la literatura como tabla de contingencia o matriz de errores."

Asimismo, estos autores establecen que los errores cometidos por el modelo de clasificación en la matriz de confusión se miden a través de los siguientes parámetros:

- Verdadero positivo (True Positive, TP): número de clasificaciones correctas en la clase positiva (P).
- Verdadero negativo (True Negative, TN): número de clasificaciones correctas en la clase negativa (N).
- Falso negativo (False Negative, FN): número de clasificaciones incorrectas de clase positiva clasificada como negativa.
- Falso positivo (False Positive, FP): número de clasificaciones incorrectas de clase negativa clasificada como positiva.

Una matriz de confusión se representa de la siguiente forma:

Ilustración 6: matriz de confusión



Fuente: Gironés Roig, J., Casas Roma, J., Minguillón Alfonso, J., & Caihuelas Quiles, R. (2017). Minería de datos: modelos y algoritmos. Barcelona: Editorial UOC.

Como se ha señalado con anterioridad, muchas de las métricas utilizadas se encuentran derivadas de la matriz de confusión mostrada anteriormente. En concreto, las métricas que se derivan de esta son: la precisión, la cobertura y la medida F.

5.4.1. Precisión.

Según Girones et al. (2017, p. 81), la precisión (PRE) mide "el rendimiento relacionado con las tasas de verdaderos positivos y negativos". La formula de la precisión es: $PRE = tp / (tp + fp)$. Por otro lado, Corso (2009, p. 8) afirma que la

precisión "mide el número de términos correctamente reconocidos respecto al total de términos predichos, sean estos verdaderos o falsos términos."

5.4.2. Cobertura.

Según Corso (2009, p. 8), recall (REC) o cobertura mide "la proporción de términos correctamente reconocidos respecto al total de términos reales, dicho de otro modo, mide en qué grado están todos los que son." La formula de la cobertura es: REC = tp / (tp + fn).

5.4.3. Medida – F.

De acuerdo con Corso (2009, p. 8), la medida F (F-Measure) sirve "para caracterizar con único valor la bondad de un clasificador o algoritmo." La formula de la medida F es*: F1 = 2 * (precision * recall) / (precision + recall).*

Por otro lado, existen métricas que no derivan directamente de la matriz de confusión pero que sí están relacionadas. Un caso es la métrica de la Curva ROC.

5.4.4. Curva ROC.

Según Girones et al. (2017, p. 82), una curva ROC mide:

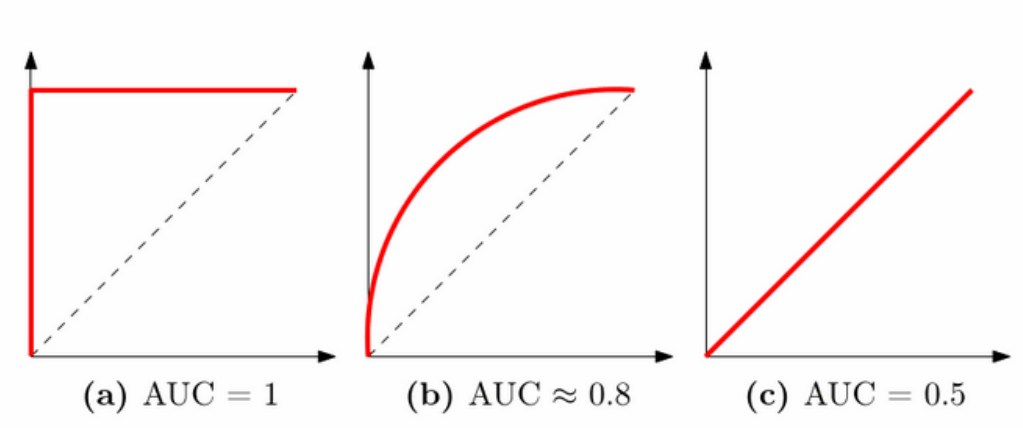
> "Una curva ROC (acrónimo de Receiver Operating Characteristic) mide el rendimiento respecto a los falsos positivos (FP) y verdaderos positivos (TP). La diagonal de la curva ROC se interpreta como un modelo generado aleatoriamente, mientras que valores inferiores se consideran peores que una estimación aleatoria de los nuevos datos."

Además, Girones et al. (2017, p. 83) aseveran que "a partir de la curva ROC es posible calcular el área debajo de la curva (AUC) que se muestra el rendimiento del modelo utilizado para la clasificación." Esta área debajo de la curva (AUC) establece los intervalos siguientes:

- [0, 5, 0, 6): Test malo
- [0, 6, 0, 75): Test regular
- [0, 75, 0, 9): Test bueno
- [0, 9, 0, 97): Test muy bueno
- [0, 97, 1): Test excelente

De forma visual, una curva ROC se representa de la siguiente forma:

Ilustración 7: curva ROC



Fuente: Gironés Roig, J., Casas Roma, J., Minguillón Alfonso, J., & Caihuelas Quiles, R. (2017). Minería de datos: modelos y algoritmos. Barcelona: Editorial UOC.

Una vez entendidas las distintas métricas para evaluar los algoritmos, algo ineludible para ejecutar estos procesos es el software. El software que se use para la aplicación de técnicas de minería de datos es una parte importante del proyecto pues, en base a las funcionalidades que pueda tener el mismo, las operaciones con los datos son más amplias y, en algunos casos, más eficientes.

5.5. Software para minería de datos

Existe una amplia gama de aplicaciones de software para minería de datos. Como en todo tipo de paquetes de software es posible diferenciar dos grandes clases: software propietario y software no propietario u open source. A continuación se hará una breve descripción de aquellas con licencia de libre (por dicha condición). Las aplicaciones destacadas son:

- R-Studio.
- RapidMiner.
- WEKA.
- KNIME.
- Orange.

5.5.1. R-Studio.

Zupan y Demšar (2008, p. 42), definen R-Studio de la siguiente forma:

> "R is a language and environment for statistical computing and graphics. Most of its computationally intensive methods are efficiently implemented in C, C++, and Fortran, and then interfaced to R, a scripting language similar to the S language originally developed at Bell Laboratories. R includes an extensive variety of techniques for statistical testing, predictive modeling, and data visualization, and has become a *de facto* standard open-source library for statistics."

5.5.2. RapidMiner.

Ristoski, Beizer y PaulHeim (2015, p. 3), describen RapidMiner de la siguiente forma:

> "RapidMiner is a data mining platform, in which data mining and analysis processes are designed from elementary building blocks, so called operators. Each operator performs a specific action on data, e.g., loading and storing data, transforming data, or inferring a model on data. The user can compose a process from operators by placing them on a canvas and wiring their input and output ports"

5.5.3. WEKA.

Zupan y Demšar (2008, p. 45), definen Weka de la siguiente forma:

> "Weka (Waikato Environment for Knowledge Analysis) is perhaps the best-known open-source machine learning and data mining environment. Advanced users can access its components through Java programming or through a command-line interface. For others, Weka provides a graphical user interface in an application called the Weka Knowledge Flow Environment featuring visual programming, and Weka Explorer providing a less flexible interface that is perhaps easier to use."

5.5.4. KNIME.

Berthold et al. (2009, p. 31) describen Knime de la siguiente forma:

> "KNIME, the Konstanz Information Miner offers a modular framework, which provides a graphical workbench for visual assembly and interactive execution of data pipelines. It features a powerful and intuitive user interface, enables easy integration of new modules or nodes, and allows for inter-active exploration of analysis results or trained models"

5.5.5. Orange.

Zupan y Demšar (2008, p. 51), definen Orange de la siguiente forma:

> "Orange is a data mining suite built using the same principles as KNIME and Weka KnowledgeFlow. In its graphical environment called Orange Canvas, the user places widgets on a canvas and connects them into a schema. Each widget performs some basic function, but unlike in KNIME with two data types - models and sets of instances - the signals passed around Orange's schemata may be of different types, and may include objects such as learners, classifiers, evaluation results, distance matrices, dendrograms, and so forth."

Una vez conocidos los principales elementos que afectan a la minería de datos (aplicaciones, procesos, algoritmos, métricas de valuación y software), es posible aplicar lo redactado en un caso práctico. Como se indicó en el objetivo principal del trabajo, el propósito es establecer la construcción de un clasificador de documentos a

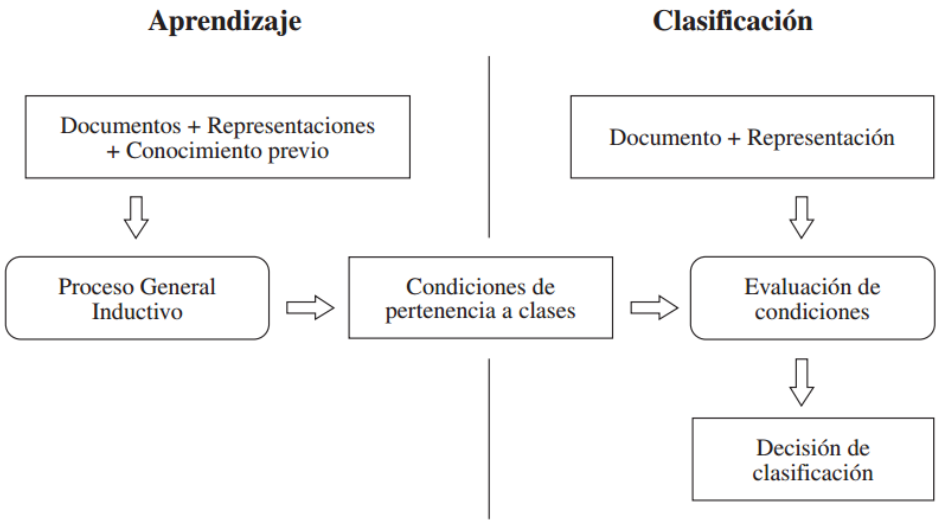
través de parte del corpus textual de Wikinoticias (véase objetivos). Por ende, los siguientes epígrafes vendrán dedicados a esta sección.

6. Clasificación Automática de Documentos con Wikinoticias. Un caso práctico

Cuando se habla de clasificación automática de textos, se habla de minería de textos o text mining. Esta subdisciplina perteneciente al campo de la minería de datos, permite extraer información relevante sobre el texto que se pretende someter a técnicas de minería de textos. A modo de ejemplo, Sun y Lim (2001, p. 521), ofrecen una definición de lo que es la Clasificación Automática de Textos: "Text classification (TC) or text categorization is the process of automatically assigning one or more predefined categories to text documents"

La Clasificación Automática de Documentos sigue el siguiente esquema:

Ilustración 8: proceso de Clasificación Automática de Documentos



Fuente: Sánchez-Jiménez, R. (2007). La documentación en el proceso de evaluación de Sistemas de Clasificación Automática. Documentación de Las Ciencias de La Información, 30, 25–44.

Si se observa el esquema, es importante destacar que se habla de documentos tanto en la fase de aprendizaje y la fase de clasificación (véase procesos en la minería de datos). Estos documentos a los que hace referencia Sánchez-Jiménez (2007) son los conjuntos de entrenamiento y de prueba. Estos conjuntos de entrenamiento y prueba se pertenecen corpus textual que se va a extraer de la fuente principal, en este caso, Wikinoticias.

Existen antecedentes de trabajos de minería de datos con proyectos Wikimedia, tal y como se señala en el trabajo de Mehdi et al. (2017), donde el corpus de Wikipedia es explotado con técnicas de minería de datos para desarrollar otros campos como el procesamiento natural del lenguaje, la construcción de ontologías o la recuperación de la información. Precisamente, en el campo de la recuperación de la información se han desarrollado diversos trabajos sobre clasificación automática de texto apoyándose en el corpus de Wikipedia (véase el apartado evaluación de Wikinoticias). A modo de

ejemplo, Mehdi et al. (2017, p. 12) conciben la Clasificación Automática de Textos en Wikipedia de la siguiente forma:

> "Text classification is a common problem in IR systems in which a classifier is trained to assign documents to appropriate classes. Studies in this section examined various methods to solve this problem benefiting from the large collection of documents available from Wikipedia. Several studies used Wikipedia knowledge base to enhance the text classication task."

En lo que a la plataforma Wikinoticias respecta, existen trabajos de minería de textos donde su corpus textual es utilizado. Un ejemplo de ello es el trabajo elaborado por Bravo-Márquez y Manriquez (2012) donde se utiliza el corpus textual de Wikinoticias para realizar labores de summarización (véase procesos en minería de datos). Sin embargo, hasta la fecha de elaboración de este trabajo, no se han identificado en la bibliografía trabajos donde se utilice el corpus textual de Wikinoticias con el objetivo de clasificar documentos. Por esta razón, se ha visto la oportunidad en este trabajo de realizar un trabajo enfocado en la Clasificación Automática de Documentos.

Para ejecutar un ejercicio de Clasificación Automática de Documentos se necesita de la ejecución de dos pasos previos:

- La constitución de un corpus textual
- La selección del software adecuado

6.1. Corpus textual de Wikinoticias

El corpus textual que se procederá a extraer de Wikinoticias se divide en dos partes:

- Corpus de entrenamiento
- Corpus de prueba

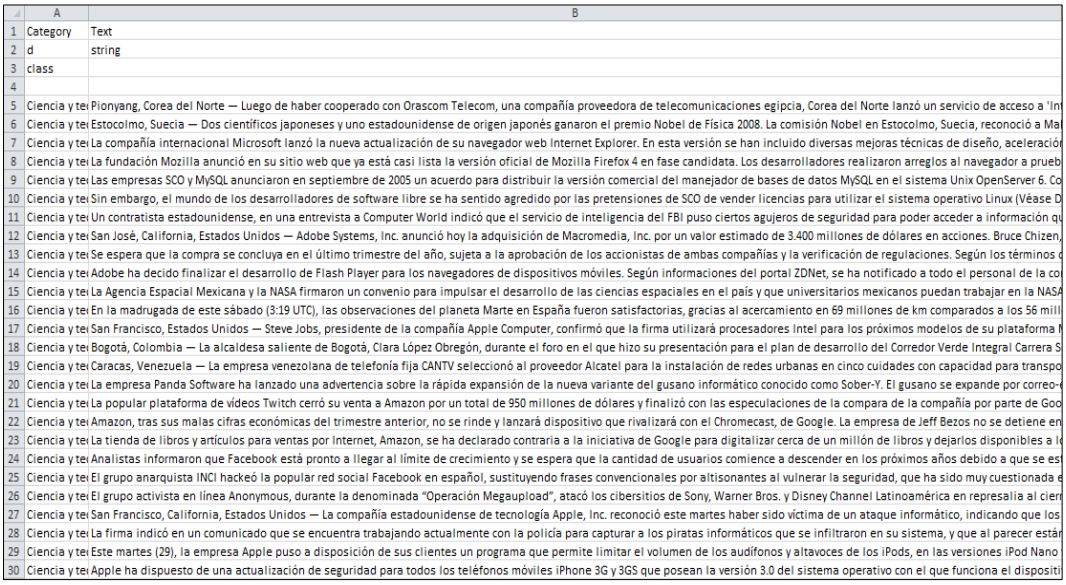
6.1.1. Corpus de entrenamiento.

El corpus de entrenamiento es el conjunto de textos que se recopilará de la fuente principal de datos, que, en este caso, es Wikinoticias. Es llamado corpus de entrenamiento porque este conjunto de textos permitirá que el algoritmo que sea aplicado pueda aprender de dicho conjunto de datos. En pocas palabras, este conjunto de datos es la fuente de alimentación del clasificador para su aprendizaje. Asimismo, utilizando la visión proporcionada por Baeza-Yates y Ribeiro-Neto (1999), este es un problema de aprendizaje es supervisado por la razón de se habla de clasificación y porque se produce una asignación previa de etiquetas al conjunto de datos (véase algoritmos y técnicas de minería de datos). Por lo tanto, para realizar esta fase, es indispensable recopilar un conjunto de documentos suficiente para entrenar el clasificador. En la práctica, esto se ha realizado de la siguiente forma:

1. Se seleccionaron cuatro categorías bien diferenciadas dentro de Wikinoticias: Ciencia y tecnología, política, salud y deportes (véase tabla de categorías de Wikinoticias). Cada una de estas categorías tendrá 200 documentos. Por lo tanto, el corpus de entrenamiento tendrá 800 documentos en total. La selección de estas noticias se realizó evitando temas repetitivos, es decir, se quiso conseguir la mayor variedad de noticias dentro de cada una de las categorías seleccionadas.
2. Una vez seleccionadas las noticias, estas fueron integradas en un bloc de notas para realizar una limpieza previa en el texto. Con ello lo que se hace es
3. Una vez limpio el texto, cada artículo es traspasado y etiquetado a un documento Excel. Este formato de hoja de tabla ha sido seleccionado por su facilidad de importación en las herramientas de minería de datos.

De forma visual, el objeto resultante es el siguiente:

Ilustración 9: corpus de entrenamiento



	A	B
1	Category	Text
2	d	string
3	class	
4		
5	Ciencia y tec	Pionyang, Corea del Norte — Luego de haber cooperado con Orascom Telecom, una compañía proveedora de telecomunicaciones egipcia, Corea del Norte lanzó un servicio de acceso a 'In
6	Ciencia y tec	Estocolmo, Suecia — Dos científicos japoneses y uno estadounidense de origen japonés ganaron el premio Nobel de Física 2008. La comisión Nobel en Estocolmo, Suecia, reconoció a Ma
7	Ciencia y tec	La compañía internacional Microsoft lanzó la nueva actualización de su navegador web Internet Explorer. En esta versión se han incluido diversas mejoras técnicas de diseño, aceleración
8	Ciencia y tec	La fundación Mozilla anunció en su sitio web que ya está casi lista la versión oficial de Mozilla Firefox 4 en fase candidata. Los desarrolladores realizaron arreglos al navegador a prueb
9	Ciencia y tec	Las empresas SCO y MySQL anunciaron en septiembre de 2005 un acuerdo para distribuir la versión comercial del manejador de bases de datos MySQL en el sistema Unix OpenServer 6. Co
10	Ciencia y tec	Sin embargo, el mundo de los desarrolladores de software libre se ha sentido agredido por las pretensiones de SCO de vender licencias para utilizar el sistema operativo Linux (Véase D
11	Ciencia y tec	Un contratista estadounidense, en una entrevista a Computer World indicó que el servicio de inteligencia del FBI puso ciertos agujeros de seguridad para poder acceder a información qu
12	Ciencia y tec	San José, California, Estados Unidos — Adobe Systems, Inc. anunció hoy la adquisición de Macromedia, Inc. por un valor estimado de 3.400 millones de dólares en acciones. Bruce Chizen,
13	Ciencia y tec	Se espera que la compra se concluya en el último trimestre del año, sujeta a la aprobación de los accionistas de ambas compañías y la verificación de regulaciones. Según los términos
14	Ciencia y tec	Adobe ha decido finalizar el desarrollo de Flash Player para los navegadores de dispositivos móviles. Según informaciones del portal ZDNet, se ha notificado a todo el personal de la co
15	Ciencia y tec	La Agencia Espacial Mexicana y la NASA firmaron un convenio para impulsar el desarrollo de las ciencias espaciales en el país y que universitarios mexicanos puedan trabajar en la NASA
16	Ciencia y tec	En la madrugada de este sábado (3:19 UTC), las observaciones del planeta Marte en España fueron satisfactorias, gracias al acercamiento en 69 millones de km comparados a los 56 mill
17	Ciencia y tec	San Francisco, Estados Unidos — Steve Jobs, presidente de la compañía Apple Computer, confirmó que la firma utilizará procesadores Intel para los próximos modelos de su plataforma M
18	Ciencia y tec	Bogotá, Colombia — La alcaldesa saliente de Bogotá, Clara López Obregón, durante el foro en el que hizo su presentación para el plan de desarrollo del Corredor Verde Integral Carrera S
19	Ciencia y tec	Caracas, Venezuela — La empresa venezolana de telefonía fija CANTV seleccionó al proveedor Alcatel para la instalación de redes urbanas en cinco cuidades con capacidad para transpo
20	Ciencia y tec	La empresa Panda Software ha lanzado una advertencia sobre la rápida expansión de la nueva variante del gusano informático conocido como Sober-Y. El gusano se expande por correo-e
21	Ciencia y tec	La popular plataforma de vídeos Twitch cerró su venta a Amazon por un total de 950 millones de dólares y finalizó con las especulaciones de la compara de la compañía por parte de Goo
22	Ciencia y tec	Amazon, tras sus malas cifras económicas del trimestre anterior, no se rinde y lanzará dispositivo que rivalizará con el Chromecast, de Google. La empresa de Jeff Bezos no se detiene en
23	Ciencia y tec	La tienda de libros y artículos para ventas por Internet, Amazon, se ha declarado contraria a la iniciativa de Google para digitalizar cerca de un millón de libros y dejarlos disponibles a l
24	Ciencia y tec	Analistas informaron que Facebook está pronto a llegar al límite de crecimiento y se espera que la cantidad de usuarios comience a descender en los próximos años debido a que se est
25	Ciencia y tec	El grupo anarquista INCI hackeó la popular red social Facebook en español, sustituyendo frases convencionales por altisonantes al vulnerar la seguridad, que ha sido muy cuestionada e
26	Ciencia y tec	El grupo activista en línea Anonymous, durante la denominada "Operación Megaupload", atacó los cibersitios de Sony, Warner Bros. y Disney Channel Latinoamérica en represalia al cier
27	Ciencia y tec	San Francisco, California, Estados Unidos — La compañía estadounidense de tecnología Apple, Inc. reconoció este martes haber sido víctima de un ataque informático, indicando que los
28	Ciencia y tec	La firma indicó en un comunicado que se encuentra trabajando actualmente con la policía para capturar a los piratas informáticos que se infiltraron en su sistema, y que al parecer están
29	Ciencia y tec	Este martes (29), la empresa Apple puso a disposición de sus clientes un programa que permite limitar el volumen de los audífonos y altavoces de los iPods, en las versiones iPod Nano
30	Ciencia y tec	Apple ha dispuesto de una actualización de seguridad para todos los teléfonos móviles iPhone 3G y 3GS que posean la versión 3.0 del sistema operativo con el que funciona el dispositi

Una parte muy importante de la elaboración del conjunto de datos para el entrenamiento es el etiquetado. Las etiquetas son la asignación a la categoría que pertenece cada texto, en este caso, ciencia y tecnología, política, salud o deporte. Para empezar, hay que saber cuál es la naturaleza de las mismas, es decir, su asignación como variables. En este caso, se trata de variables nominales o cualitativas pues solo nos dan información sobre la misma etiqueta. A propósito de esto, Han, Kamber y Pei (2012, p. 41) realizan el siguiente apunte: "Nominal means "relating to names." The values of a nominal attribute are symbols or names of things. Each value represents some kind of category, code, or state, and so nominal attributes are also referred to as categorical"

6.1.2. Corpus de prueba.

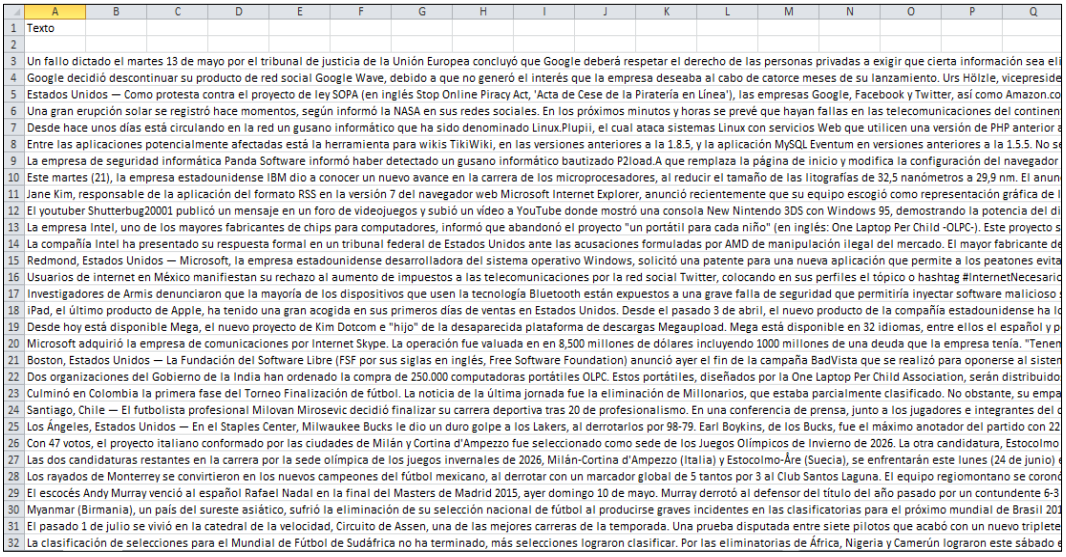
El corpus de prueba es el conjunto de textos que se recopila para poner a prueba la capacidad de acierto del clasificador elaborado. De la misma forma que el corpus de

entrenamiento, este será extraído también de Wikinoticias con la salvedad de que este no será etiquetado. El proceso de recopilación de textos para el corpus de prueba es el mismo que el anterior con ciertas modificaciones. Por tanto, el proceso de selección de textos vendrá dado de la siguiente forma:

1. Se seleccionarán 20 noticias al azar de cada una de las categorías indicadas anteriormente (Ciencia y tecnología, política, salud y deporte).
2. Una vez seleccionadas, serán depositadas en un bloc de notas para realizar una limpieza en el texto.
3. Finalmente, estas serán trasladadas a un archivo Excel que permitirá su lectura en el software especializado en minería de datos

De forma ilustrativa, esto figura de la siguiente forma:

Ilustración 10: corpus de prueba



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Texto																
2																	
3	Un fallo dictado el martes 13 de mayo por el tribunal de justicia de la Unión Europea concluyó que Google deberá respetar el derecho de las personas privadas a exigir que cierta información sea eli																
4	Google decidió descontinuar su producto de red social Google Wave, debido a que no generó el interés que la empresa deseaba al cabo de catorce meses de su lanzamiento. Urs Hölzle, vicepreside																
5	Estados Unidos — Como protesta contra el proyecto de ley SOPA (en inglés Stop Online Piracy Act, 'Acta de Cese de la Piratería en Línea'), las empresas Google, Facebook y Twitter, así como Amazon.co																
6	Una gran erupción solar se registró hace momentos, según informó la NASA en sus redes sociales. En los próximos minutos y horas se prevé que hayan fallas en las telecomunicaciones del continen																
7	Desde hace unos días está circulando en la red un gusano informático que ha sido denominado Linux.Plupii, el cual ataca sistemas Linux con servicios Web que utilicen una versión de PHP anterior a																
8	Entre las aplicaciones potencialmente afectadas está la herramienta para wikis TikiWiki, en las versiones anteriores a la 1.8.5, y la aplicación MySQL Eventum en versiones anteriores a la 1.5.5. No se																
9	La empresa de seguridad informática Panda Software informó haber detectado un gusano informático bautizado P2load.A que remplaza la página de inicio y modifica la configuración del navegador																
10	Este martes (21), la empresa estadounidense IBM dio a conocer un nuevo avance en la carrera de los microprocesadores, al reducir el tamaño de las litografías de 32,5 nanómetros a 29,9 nm. El anun																
11	Jane Kim, responsable de la aplicación del formato RSS en la versión 7 del navegador web Microsoft Internet Explorer, anunció recientemente que su equipo escogió como representación gráfica de l																
12	El youtuber Shutterbug20001 publicó un mensaje en un foro de videojuegos y subió un vídeo a YouTube donde mostró una consola New Nintendo 3DS con Windows 95, demostrando la potencia del di																
13	La empresa Intel, uno de los mayores fabricantes de chips para computadores, informó que abandonó el proyecto "un portátil para cada niño" (en inglés: One Laptop Per Child -OLPC-). Este proyecto s																
14	La compañía Intel ha presentado su respuesta formal en un tribunal federal de Estados Unidos ante las acusaciones formuladas por AMD de manipulación ilegal del mercado. El mayor fabricante de																
15	Redmond, Estados Unidos — Microsoft, la empresa estadounidense desarrolladora del sistema operativo Windows, solicitó una patente para una nueva aplicación que permite a los peatones evita																
16	Usuarios de internet en México manifiestan su rechazo al aumento de impuestos a las telecomunicaciones por la red social Twitter, colocando en sus perfiles el tópico o hashtag #InternetNecesaric																
17	Investigadores de Armis denunciaron que la mayoría de los dispositivos que usen la tecnología Bluetooth están expuestos a una grave falla de seguridad que permitiría inyectar software malicioso																
18	iPad, el último producto de Apple, ha tenido una gran acogida en sus primeros días de ventas en Estados Unidos. Desde el pasado 3 de abril, el nuevo producto de la compañía estadounidense ha lo																
19	Desde hoy está disponible Mega, el nuevo proyecto de Kim Dotcom e "hijo" de la desaparecida plataforma de descargas Megaupload. Mega está disponible en 32 idiomas, entre ellos el español y p																
20	Microsoft adquirió la empresa de comunicaciones por Internet Skype. La operación fue valuada en en 8,500 millones de dólares incluyendo 1000 millones de una deuda que la empresa tenía. "Tenen																
21	Boston, Estados Unidos — La Fundación del Software Libre (FSF por sus siglas en inglés, Free Software Foundation) anunció ayer el fin de la campaña BadVista que se realizó para oponerse al sisten																
22	Dos organizaciones del Gobierno de la India han ordenado la compra de 250.000 computadoras portátiles OLPC. Estos portátiles, diseñados por la One Laptop Per Child Association, serán distribuido																
23	Culminó en Colombia la primera fase del Torneo Finalización de fútbol. La noticia de la última jornada fue la eliminación de Millonarios, que estaba parcialmente clasificado. No obstante, su empa																
24	Santiago, Chile — El futbolista profesional Milovan Mirosevic decidió finalizar su carrera deportiva tras 20 de profesionalismo. En una conferencia de prensa, junto a los jugadores e integrantes del c																
25	Los Ángeles, Estados Unidos — En el Staples Center, Milwaukee Bucks le dio un duro golpe a los Lakers, al derrotarlos por 98-79. Earl Boykins, de los Bucks, fue el máximo anotador del partido con 22																
26	Con 47 votos, el proyecto italiano conformado por las ciudades de Milán y Cortina d'Ampezzo fue seleccionado como sede de los Juegos Olímpicos de Invierno de 2026. La otra candidatura, Estocolmo																
27	Las dos candidaturas restantes en la carrera por la sede olímpica de los juegos invernales de 2026, Milán-Cortina d'Ampezzo (Italia) y Estocolmo-Åre (Suecia), se enfrentarán este lunes (24 de junio) e																
28	Los rayados de Monterrey se convirtieron en los nuevos campeones del fútbol mexicano, al derrotar con un marcador global de 5 tantos por 3 al Club Santos Laguna. El equipo regiomontano se coronó																
29	El escocés Andy Murray venció al español Rafael Nadal en la final del Masters de Madrid 2015, ayer domingo 10 de mayo. Murray derrotó al defensor del título del año pasado por un contundente 6-3																
30	Myanmar (Birmania), un país del sureste asiático, sufrió la eliminación de su selección nacional de fútbol al producirse graves incidentes en las clasificatorias para el próximo mundial de Brasil 201																
31	El pasado 1 de julio se vivió en la catedral de la velocidad, Circuito de Assen, una de las mejores carreras de la temporada. Una prueba disputada entre siete pilotos que acabó con un nuevo triplete																
32	La clasificación de selecciones para el Mundial de Fútbol de Sudáfrica no ha terminado, más selecciones lograron clasificar. Por las eliminatorias de África, Nigeria y Camerún lograron este sábado e																

Para comprender mejor la elaboración de ambos corpus, es necesario mostrar la composición del corpus total a través de dos tablas:

Tabla 9: composición del corpus de entrenamiento

Categoría	**Número de artículos**
Ciencia y tecnología	200
Política	200
Salud	200
Deporte	200
Número total de artículos:	**800**

Tabla 10: composición del corpus de prueba

Categoría	Número de artículos
Ciencia y tecnología (sin etiquetar)	20
Política (sin etiquetar)	20
Salud (sin etiquetar)	20
Deporte (sin etiquetar)	20
Número total de artículos:	**80**

Una vez elaborado el corpus de Wikinoticias, el siguiente paso es seleccionar la herramienta adecuada para ejecutar el proceso de minería de datos.

6.2. Selección de software

La selección del software adecuado para el caso aplicado es una fase indispensable para la ejecución del mismo. Respecto a esto, Brindha, Prabha y Sukumaran (2016, p. 1) realizan el siguiente apunte: "The tool could be a text mining system that has the capability to analyse giant quantities of natural language text and detects lexical and linguistic usage patterns in an attempt to meaningful and helpful information." Como se ha podido observar en epígrafes anteriores, existen diversas herramientas especializadas en minería de datos. Por suerte, muchas de estas herramientas son software libre (véase software para minería de datos). Además, es posible resaltar que estos productos no propietarios ofrecen muy buenas especificaciones y que la variedad de los mismos es positiva para los consumidores de este tipo de software. Existen trabajos como el de Al-Odan y Al-Daraiseh (2015) que realizan un estudio comparativo sobre herramientas de software libre que tratan con minería de datos y que, por ende, ayudan a seleccionar la herramienta que más se adecúe a las necesidades del consumidor. De manera visual, los autores representan este estudio comparativo mediante la tabla siguiente:

Ilustración 11: evaluación de herramientas para minería de datos

	RStudio	RapidMiner	WEKA	KNIME	Orange
Intuitiveness	5.88	7.88	6.41	7.94	**8.06**
Consistency	7.24	7.47	5.94	**7.82**	7.65
Navigation	6.82	**7.94**	6.29	**7.94**	7.59
Usability	6.47	**8.35**	6.06	8	8
Installation Manual	3.76	**8.89**	6.12	8.41	8.35
Configuration Guide	4.18	**8.76**	5.94	8.47	8.06
Troubleshooting Guide	6.41	**8.35**	6.24	8.06	7.47
User Tutorials	6.12	**8.71**	6.47	8.35	8.18

Fuente: Al-Odan, H. A., & Al-Daraiseh, A. A. (2015). Open Source Data Mining tools. In 2015 International Conference on Electrical and Information Technologies (ICEIT) (pp. 369–374). Marrakech, Morocco: IEEE.

Analizando los resultados del estudio en base a los métodos de evaluación propuestos, RapidMiner es el software que mejores puntuaciones tiene. A este le siguen KNIME y Orange. Precisamente, se ha hablado de las herramientas que figuran en la tabla en epígrafes anteriores (véase software para minería de datos).

A pesar de que RapidMiner y KNIME han tenido muy buenos resultados en estudio, en el caso práctico se usará Orange. Según el estudio, es el software más intuitivo seleccionado por los usuarios. Esto provoca que la balanza se decante por esta aplicación ya que, al no ser un usuario experto en el uso de esta clase de herramientas, el valor añadido de que la aplicación sea intuitiva ahorra tiempo de aprendizaje y, por tanto, de ejecución del proyecto.

Por otro lado y de forma positiva, la tabla indica que Orange tiene muy buenas especificaciones con respecto a otras herramientas pues en todos los métodos de evaluación propuestos por los autores, en ninguno ofrece resultados por debajo del 7. Por tanto, en este trabajo se utilizara el software Orange para el caso práctico.

6.2.1. Orange.

Orange Canvas es una herramienta software libre basada en lenguaje Python que permite realizar tareas de data mining y que es capaz de aplicar técnicas de machine learning, entre otras cosas. Destacada por su vistosa e intuitiva interfaz basada en la combinación de diversos widgets sobre un espacio de trabajo y desarrollada por la Universidad de Ljubljana en su Laboratorio de Bioinformática, es una de las herramientas más conocidas del panorama open source en cuanto a software de minería de datos (Naik y Samant, 2016). Esta herramienta está diseñada tanto para usuarios experimentados como para usuarios que comienzan a introducirse en el mundo de la minería de datos (véase selección de software). Demšar, Zupan, Leban y Curk (2004, p. 1) señalan lo siguiente en referencia a Orange Canvas:

> "It is intended for both experienced users and researchers in machine learning who want to write Python scripts to prototype new algorithms while reusing as much of the code as possible, and for those just entering the field who can enjoy in the powerful while easy-to-use visual programming environment."

Para mostrar todas las posibilidades que ofrece Orange, se mostrará parte de la tabla que destacan Al-Odan y Al-Daraiseh (2015) en su análisis de las propiedades del software:

Tabla 11: especificaciones de Orange

		Orange Canvas
Platform Support	Windows	✓
	Mac	✓
	Lynux	✓
Interface	Text	
	GUI	
	Interactive GUI	✓
Installation Process	Single Package	✓
	Multi Package	
	Online/Server Application	
	Flat Files	
	Developer Version	✓
Data Sources	MS Access	
	MS Excel	✓
	MySQL	✓
	ARFF	✓
	CSV	✓
Supported Algorithms	Decision Trees	✓
	Linear/Statistical	✓
	Bayes	✓
	Neural Networks	✓
	K Means	✓
	Nearest Neigbor	✓
Output	Bar Charts	✓
	Pie Charts	✓
	Scatter Plots	✓
	Classification Trees	✓

Fuente: Al-Odan, H. A., & Al-Daraiseh, A. A. (2015). Open Source Data Mining tools. In 2015 International Conference on Electrical and Information Technologies (ICEIT) (pp. 369–374). Marrakech, Morocco: IEEE.

Visualizando la tabla anterior, Orange cumple con las expectativas, ofreciendo todas las posibilidades en cuanto a algoritmos, tanto supervisados como no supervisados (véase algoritmos y técnicas en minería de datos). Asimismo, su guía interactiva facilita el aprendizaje para usuarios que no son expertos en la materia. En definitiva, Orange tiene todo lo necesario para abordar el caso práctico.

7. Caso aplicado con Orange

Una vez seleccionada la herramienta, es posible proceder a extraer la información relevante de los datos o, lo que es lo mismo, iniciar el proceso de knowledge discovery from data (KDD). Para resolver este problema de clasificación, se seguirá el proceso de minería de datos propuesto por Molina (2002, p. 4) que establece 4 fases:

1. Determinación de los objetivos.
2. Preprocesamiento de los datos.
3. Determinación del modelo.
4. Análisis de los resultados.

7.1. Determinación de los objetivos

La determinación de los objetivos antes de proceder a realizar labores de minería de datos es un paso crucial. Los objetivos son los siguientes:

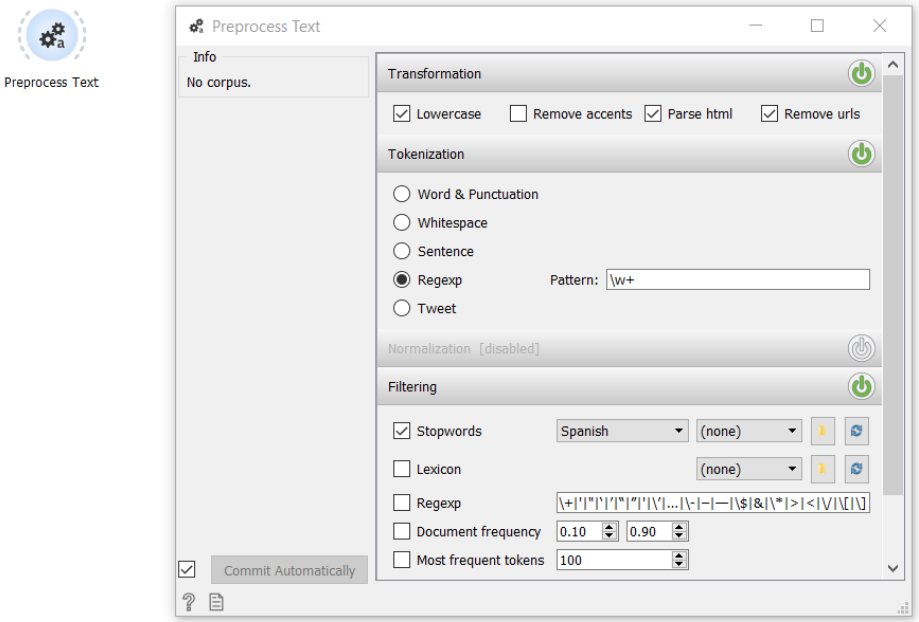
- Objetivo general: crear un clasificador que sea capaz de asignar la categoría correcta a un conjunto de 80 documentos que se encuentran sin etiquetar.
- Objetivo específico: comparar y seleccionar el algoritmo que mejor se ajuste al problema de clasificación presentado anteriormente.

Una vez establecidos los objetivos, es posible empezar a trabajar con los datos. El siguiente paso es la eliminación de palabras vacías y caracteres poco relevantes del corpus a través del preprocesamiento. Esto permitirá que el corpus textual quede "limpio" en mayor o menor medida.

7.2. Preprocesamiento de los datos

La parte del preprocesamiento está enfocada a la eliminación de ruido en los datos. Este ruido normalmente está identificado con palabras vacías o stopwords, caracteres y símbolos. Asimismo, el preprocesamiento de datos está relacionado con la normalización de los mismos. A este proceso también se le llama limpieza de datos. En Orange, el preprocesamiento se visualiza de la siguiente forma:

Ilustración 12: preprocesamiento



Una buena manera de saber si se ha realizado bien la labor de preprocesamiento es utilizando las bolsas de palabras. Esta herramienta permite ver las palabras destacadas de un corpus en base a su frecuencia de aparición en el mismo.

7.2.1. Bolsas de palabras.

En base a lo que Layton, Watters y Dazeley (2012) proponen, las bolsas de palabras (Bag of Words) están clasificadas dentro de las formas dinámicas para representar documentos y sirven para detectar los términos más representativos de un conjunto de documentos. Una vez realizada la labor de preprocesamiento, con las bolsas de palabras es posible observar los términos que más se repiten en el conjunto de documentos que va a ser utilizado para entrenar el modelo. La forma más sencilla de representar una bolsa de palabras es a través de una nube de palabras (Word Cloud). En Orange, una nube de palabras se representa de la siguiente forma:

Ilustración 13: bolsa de palabras representada por una nube de palabras

Fuente: elaboración propia

Una vez se han representado las palabras en la nube de palabras, es posible observar que las palabras vacías y los caracteres han sido eliminados. Por lo tanto, quedarían solamente aquellas palabras relevantes del corpus. Con esto finalizado, es posible proceder al paso siguiente: la determinación del modelo.

7.3. Determinación del modelo

Para determinar el modelo que mejor se ajusta a las necesidades del problema, en primera instancia hay que tener claro qué clase de problema se identifica, es decir, hay que saber si se trata de un problema donde se tengan que utilizar métodos supervisados o métodos no supervisados. Al tratarse, en este caso, de una labor de clasificación, se optara por el uso de métodos supervisados. Además, observando el contexto del problema, hay que destacar que se tienen dos conjuntos de datos bien diferenciados: uno para entrenamiento y otro para poner a prueba el modelo. El conjunto de datos de entrenamiento viene etiquetado, condición sine qua non para resolver el problema de la clasificación.

Existen numerosos trabajos que abordan la selección del método adecuado o con mejores resultados para clasificar textos. Para ello, normalmente los autores suelen realizar comparativas entre los algoritmos destacados para realizar la clasificación. Precisamente, Calderón et al. (2016, p. 629), aportan lo siguiente sobre los algoritmos supervisados:

> “Para el aprendizaje supervisado, modelos como el de predicción basado en máquinas de vectores soporte (SVM) que pasan nuestros datos a un espacio multidimensional, permiten crear algoritmos potentes a partir de datos existentes

> (un ejemplo: conjuntos de noticias ya separadas por tema) para crear patrones que permitan categorizar automáticamente nuevos conjuntos de textos. Esto es muy útil para la clasificación automática de noticias."

A pesar de que Calderón et al. (2016) hacen hincapié en la técnica de maquinas de vectores soporte (Support Vector Machines), no es la única aplicable. En el apartado de algoritmos y técnicas de minería de datos, es posible ver que existe más de un método para aplicar. En Orange, las técnicas a aplicar son:

- Árboles de decisión
- Naïve Bayes
- K-vecinos más cercanos

7.3.1. Árboles de decisión.

Según explican Goddard et al. (1995, p. 3), el algoritmo de árboles de decisión funciona de la siguiente forma:

> "El algoritmo utilizado genera AD binarios, es decir, divide el conjunto de ejemplos en cada nodo en dos partes. La selección del atributo para realizar esta partición depende de la información, en términos de la clasificación de los grupos de salida, contenida en cada atributo. El atributo con más información es seleccionado para realizar la partición."

Por otro lado, Girones et al. (2017, p. 209) aseveran lo siguiente:

> "Los árboles de decisión (Decision Trees) son uno de los modelos de minería de datos más comunes y estudiados, y no precisamente por su capacidad predictiva, superada generalmente por otros modelos más complejos, sino por su alta capacidad explicativa y la facilidad para interpretar el modelo generado."

7.3.2. Naïve Bayes.

Para Gironés et al. (2017, p. 229) "los métodos probabilísticos, o también llamados métodos estadísticos suelen estimar un conjunto de parámetros probabilísticos, que expresan la probabilidad condicionada de cada clase dadas las propiedades de un ejemplo (descrito en forma de atributos)". Entre los métodos probabilísticos destacados, se encuentra Naïve Bayes, algoritmo basado en el teorema de Bayes. En cuanto a su relevancia, estos autores confirman lo siguiente:

> "El algoritmo de clasificación de Naïve Bayes se basa en el concepto de probabilidad condicional y busca maximizar la verosimilitud del modelo, es decir, otorgar mayor importancia a aquellos eventos que son realmente relevantes en el juego de datos."

7.3.3. K vecinos más cercanos.

Girones et al. (2017, p. 135) definen el algoritmo k- vecinos más cercanos de la siguiente forma:

> "El k-NN o «k vecinos más cercanos» (en inglés, k nearest neighbours) es un algoritmo de aprendizaje supervisado de clasificación, de modo que a partir de un juego de datos de entrenamiento su objetivo será clasificar correctamente todas las instancias nuevas. El juego de datos típico de este tipo de algoritmos está formado por varios atributos descriptivos y un solo atributo objetivo, también llamado clase."

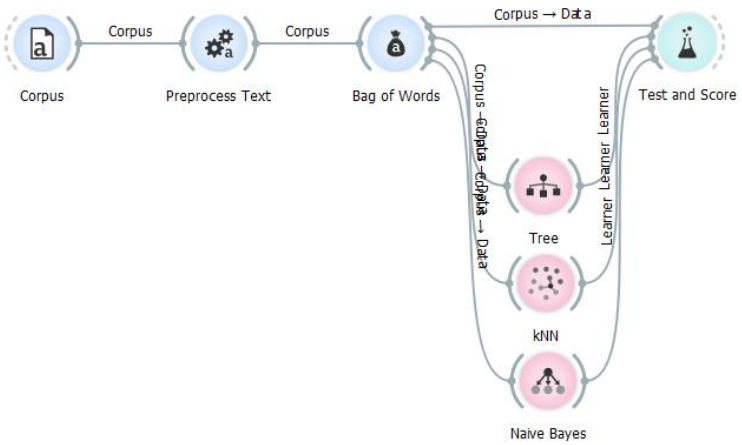
7.3.4. Prueba de los algoritmos.

Una vez conocidos los algoritmos que se van a aplicar, es posible proceder ejecutarlos. Para poner a prueba los algoritmos en Orange, se necesita:

- Conjunto de datos para entrenamiento
- Preprocesamiento
- Bolsa de palabras
- Algoritmos supervisados o predictivos
    - Árboles de decisión
    - Naïve Bayes
    - K-vecinos más cercanos

Para elaborar el modelo de clasificación adecuado, hay que ensamblar todos los componentes necesarios. En Orange, el ensamblaje de todos los componentes se realiza en el espacio de trabajo. En el software esto se encuentra representado de la siguiente manera:

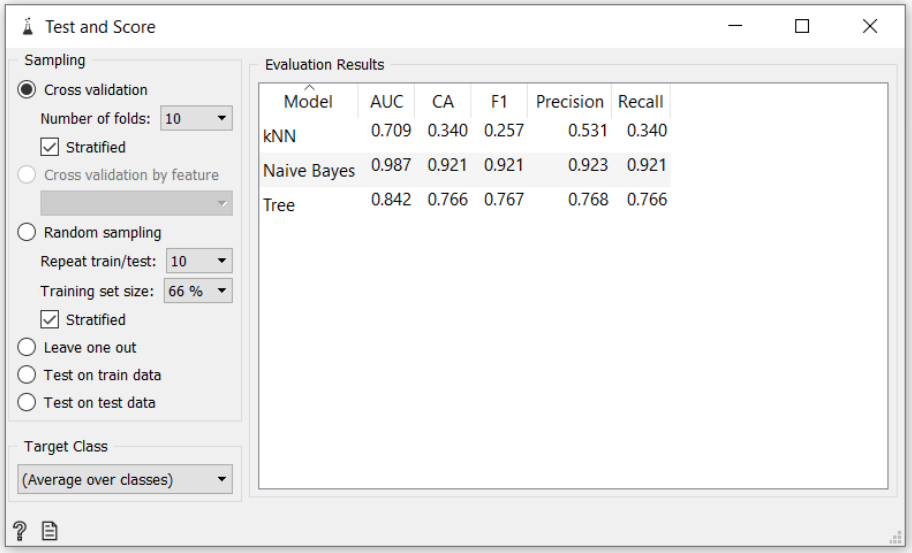
Ilustración 14: ensamblaje de widgets en para la evaluación



Fuente: elaboración propia

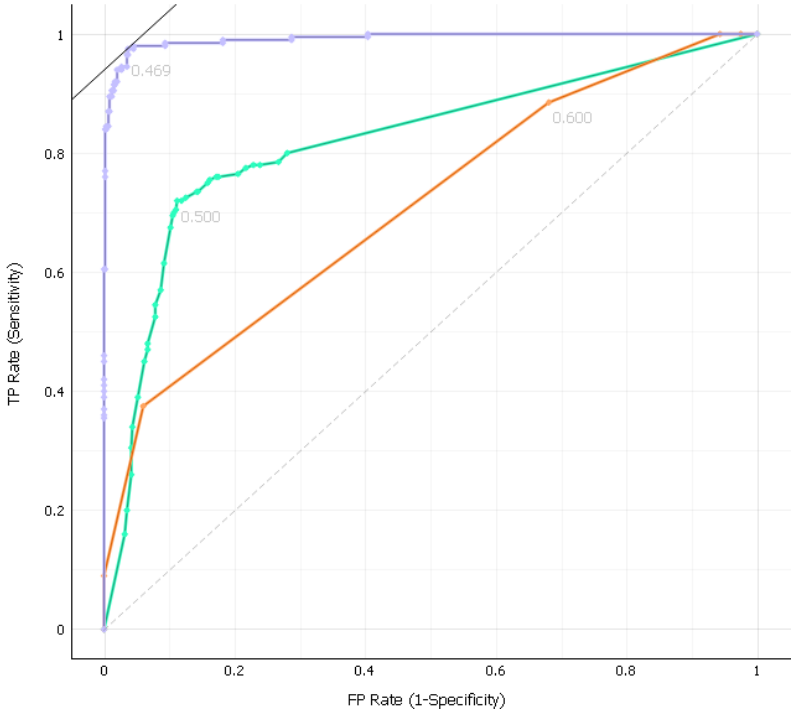
Una vez integrados todos los elementos, el resultado ofrecido por Orange es el siguiente:

Ilustración 15: resultados de los algoritmos



Analizando los resultados manifestados por la herramienta y asumiendo la condición de que sea 1,0 el mejor valor y 0,0 el peor en las distintas métricas utilizadas (véase métricas de evaluación de algoritmos), se seleccionará el algoritmo que obtenga el valor más próximo a 1. Observando los métodos de evaluación, los resultados revelan que el algoritmo que mejor resultados ofrece es Naïve Bayes. En comparación con k-vecinos más cercanos y los árboles de decisión, Naïve Bayes es el que más se acerca a 1,0 en todas las métricas, obteniendo cifras superiores a 0,9 en cada una de ellas. Por consiguiente y, con los resultados obtenidos en la red bayesiana, es posible determinar que el clasificador aquí planteado ofrece muy buenas condiciones para el solventar el problema de clasificación aquí planteado. Para representar esto de forma visual, es posible utilizar el análisis ROC para poder evaluar mejor los modelos (véase curva ROC). En Orange, esto se muestra de la siguiente manera:

Gráfico 1: análisis de la curva ROC



Fuente: elaboración propia

Acorde con los intervalos establecidos sobre el área bajo la curva del grafico (AUC), es posible identificar tres tipos de test: un test regular, un test bueno y un test excelente. Siguiendo el grafico en base a los colores, la silueta de color morado pertenece a Naïve Bayes, siendo este el test excelente, la silueta de color verde pertenece a los árboles de decisión siendo este el test bueno y, por último, la silueta color marrón pertenece al algoritmo k-vecinos más cercanos siendo este el test regular. Por tanto, la evaluación del test estaría reflejada de la siguiente forma:

Tabla 12: comparación de algoritmos con los intervalos AUC

Test utilizado	Resultado del test	Intervalo AUC	Valoración
Naïve Bayes	0,987	[0, 97, 1)	Excelente
Árboles de decisión	0,842	[0, 75, 0, 9)	Bueno
K-vecinos	0,709	[0, 6, 0, 75)	Regular

En vista a los resultados ofrecidos por las métricas, se hará uso de la técnica Naïve Bayes para poner a prueba el clasificador. Esta fase se corresponde con el análisis de resultados, la última establecida por Molina (2002, p. 4).

7.4. Análisis de los resultados

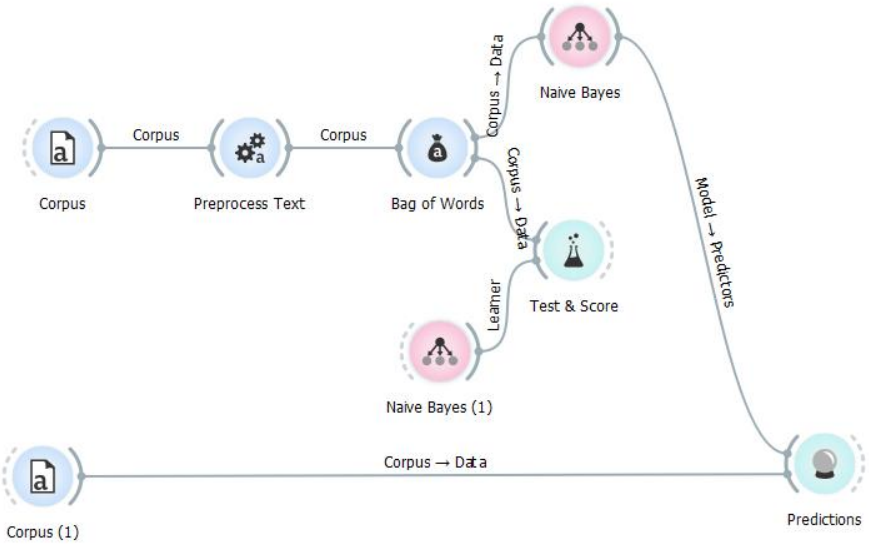
El análisis de resultados es la fase que responde al problema de clasificación planteado al principio del trabajo. Para ello, es necesario ensamblar todo el modelo que va a poner a prueba el conjunto de datos para verificar si el clasificador funciona. Tras

la selección del algoritmo adecuado, los componentes que tienen que figurar en el espacio de trabajo del software son:

- Conjunto de datos para entrenamiento
- Preprocesamiento
- Bolsa de palabras
- Algoritmos supervisados o predictivos
    - Naïve Bayes
- Conjunto de datos para evaluar

Al igual que en otros epígrafes, se necesita ensamblar todos los componentes necesarios del modelo para ponerlo a prueba. En Orange, esto queda representado de la siguiente forma:

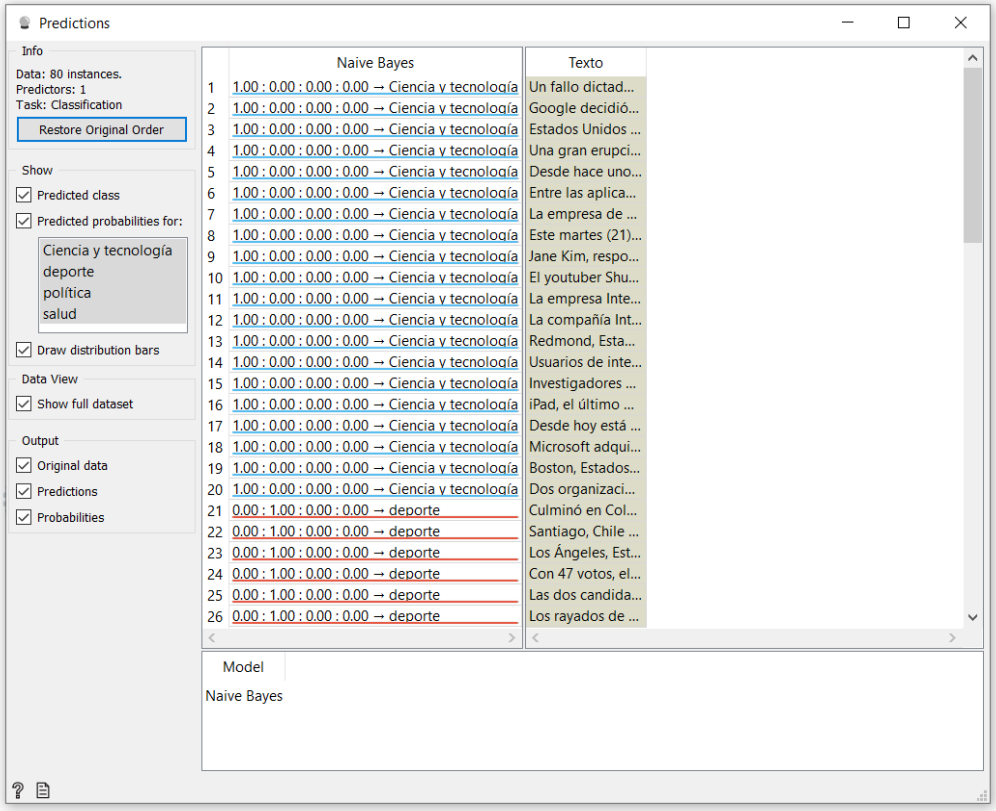
Ilustración 16: desarrollo del modelo con Naïve Bayes



Fuente: elaboración propia

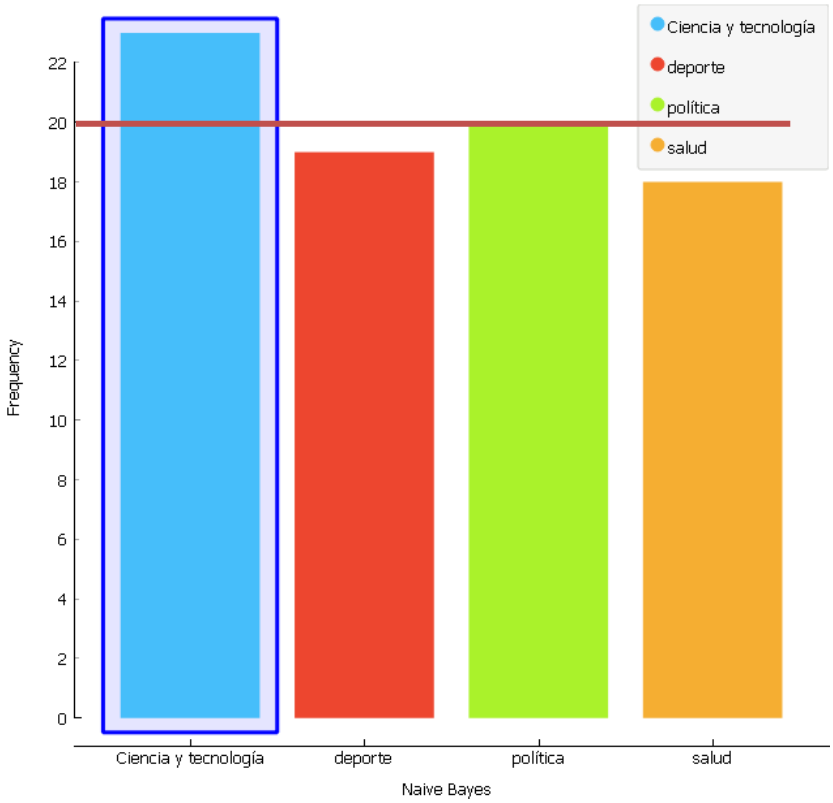
Una vez obtenido el esquema del modelo final, solo queda proceder a mostrar los resultados del test. Para mostrar los resultados hay que observar las predicciones del modelo Naïve Bayes sobre el conjunto de datos de evaluación. Esto en Orange queda ilustrado de la siguiente forma:

Ilustración 17: resultados con Naïve Bayes



A la vista de los resultados ofrecidos, es posible determinar que el algoritmo aplicado para clasificar los documentos es válido para alcanzar el objetivo principal del ejercicio. Gran parte de la bibliografía encontrada asevera que el clasificador bayesiano obtiene muy buenos resultados en comparación con otros algoritmos y, en este trabajo, eso queda demostrado. A pesar de tener un porcentaje alto de acierto, existe un ligero margen de mejora pues ha clasificado de forma incorrecta 3 de 80 textos recogidos. Para comprobar esto, el corpus de prueba fue etiquetado previamente para saber con certeza dónde es que falla el clasificador aplicado. Esto es demostrable con la gráfica siguiente:

Gráfico 2: errores cometidos con Naïve Bayes



Fuente: elaboración propia

Es visible que en la categoría ciencia y tecnología hay más documentos de los que tendría que haber. La franja roja que está situada en la parte superior del grafico marca el límite de documentos por categoría. Por ende, es observable que categorías como política y salud se encuentran carentes de documentos (los cuales fueron asignados a ciencia y tecnología con un error de 3. Sin embargo, hay que tener en cuenta que el corpus de entrenamiento no era demasiado grande. Probablemente, si el corpus de entrenamiento fuera de 1000-1500 documentos los resultados serían aún mejores que los ofrecidos actualmente.

Por otro lado, es posible destacar que el algoritmo no ha sufrido ningún tipo de sobre entrenamiento del modelo. Este efecto, también llamado sobreajuste, ocurre cuando el modelo "acabe respondiendo estrictamente a las propiedades del juego de datos de entrenamiento y que sea incapaz de extrapolarse con niveles de acierto adecuados a otros juegos de datos que puedan aparecer en un futuro." (Gironés et al., 2017, p. 72).

Por otra parte, cabe destacar que la elaboración del corpus de entrenamiento presentado en el trabajo no está elaborado por un conjunto de profesionales que se han preocupado de establecer de forma correcta las categorías ya que, la asignación de las mismas, está realizada por aquellos usuarios que publican noticias en Wikinoticias. Esto lleva a otro punto interesante que ya señalaba Sánchez-Jiménez en 2007 que es la preocupación por realizar conjuntos de entrenamiento que cuiden su análisis documental. A raíz de esto, existe mucho benchmarking en base a la oferta de grandes

corpus que sirven para entrenar pero no se tiene la completa certeza de que las categorías propuestas se encuentren bien asignadas. En referencia a lo escrito previamente Sánchez-Jiménez (2007, p. 40) realiza el siguiente apunte:

> "Desde un punto de vista objetivo la indización por descriptores y la indización por materias son tareas distintas. Si esta distinción no se tiene en cuenta a la hora de establecer corpus por los que evaluar los resultados de los sistemas de clasificación automática, no podemos concluir que dichos corpus sirvan para evaluar la calidad de los resultados de clasificación."

Por tanto, la elaboración de los conjuntos de entrenamiento para evaluar los sistemas de clasificación automática deberían de ser elaborados por profesionales entendidos en materia de indización. Esto significaría una mejora significante en el proceso de aprendizaje y, por ende, en el ejercicio de la Clasificación Automática de Documentos.

8. Conclusiones

1. Se ha podido comprobar que con una muestra de 800 documentos es posible elaborar un clasificador que cumple con las exigencias requeridas en el presente trabajo. Existe numerosa bibliografía sobre la clasificación automática de documentos y, por lo que se ha podido observar, normalmente un modelo de clasificación consigue un buen porcentaje de clasificación a partir de 1000 documentos.
2. La simplificación y difusión de las herramientas para minería de datos ha posibilitado que personas que no son necesariamente expertas puedan abordar ciertos estudios. Como reflejo de esto, la orientación del software Orange en el campo de la minería de datos, ha posibilitado la elaboración de este trabajo.
3. Dando la razón a los estudios sobre clasificación automática de documentos, el algoritmo Naïve Bayes trabaja muy bien con pocos datos y, a pesar de no ser una técnica novedosa precisamente, sus resultados son más que aceptables.
4. La plataforma de noticias de agencia libres, Wikinoticias, puede ser utilizada como fuente de datos para desarrollar experimentos sobre clasificación automática de documentos o, por otro lado, para solventar problemas de extracción de tópicos. Sus artículos revisados y su condición de uso libre, posibilitan que sea una fuente de datos propicia para la creación de corpus.
5. Con las exigencias del mercado actual, un gestor de documentos podría implementar ciertas técnicas de minería de datos para agilizar su trabajo, tal y como se expone en el presente trabajo.

9. Bibliografía

Abolhosen, N. Y. (2017). *Nuevos entornos, nueva carne: Reconfiguración y personalización tecnológica de la cultura*. Guadalajara, Jalisco: ITESO.

Al-Odan, H. A., & Al-Daraiseh, A. A. (2015). Open Source Data Mining tools. In *2015 International Conference on Electrical and Information Technologies (ICEIT)* (pp. 369–374). Marrakech, Morocco: IEEE. doi: https://doi.org/10.1109/EITech.2015.7162956

Arcila-Calderón, C., Barbosa-Caro, E., & Cabezuelo-Lorenzo, F. (2016). Técnicas Big Data: análisis de textos a gran escala para la investigación científica y periodística. *El Profesional de La Información*, *25*(4), 623. doi: https://doi.org/10.3145/epi.2016.jul.12

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval.* New York: ACM Press.

Berners-Lee, T., Weitzner, D. J., Hall, W., O'Hara, K., Shadbolt, N., & Hendler, J. A. (2006). A Framework for Web Science. *Foundations and Trends® in Web Science*, *1*(1), 1–130. doi: https://doi.org/10.1561/1800000001

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., … Wiswedel, B. (2009). KNIME - the Konstanz information miner. *ACM SIGKDD Explorations Newsletter*, *11*(1), 26–31. doi: https://doi.org/10.1145/1656274.1656280

Bravo-Marquez, F., & Manriquez, M. (2012). A zipf-like distant supervision approach for multi-document summarization using wikinews articles. In L. Calderón Benavides, C. González-Caro, E. Chávez, & N. Ziviani (Eds.), *String Processing and Information Retrieval (SPIRE)* (Vol. 7608, pp. 143–154). Cartagena de Indias, Colombia: Springer, Berlin, Heidelberg. doi: https://doi.org/10.1007/978-3-642-34109-0-15

Brindha, S., Prabha, K., & Sukumaran, S. (2016). A survey on classification techniques for text mining. In *2016 3rd International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 1–5). Coimbatore, India: IEEE. doi: https://doi.org/10.1109/ICACCS.2016.7586371

Bruns, A. (2006). Wikinews: The Next Generation of Online News? *Scan: Journal of Media Arts Culture*, *3*(1), 1–14. Retrieved from: https://eprints.qut.edu.au/4862/

Casas Roma, J., Nin Guerrero, J., & Julbe, F. (2019). *Big data : análisis de datos en entornos masivos*. Barcelona: Editorial UOC.

Codina Bonilla, L., Marcos, M. C., & Pedraza, R. (2009). *Web semántica y sistemas de información documental*. Gijón: Trea.

Corso, C. L. (2009). Aplicación de algoritmos de clasificación supervisada usando. In *Congreso Nacional de Información y Telecomunicaciones (CNIT)* (pp. 1–11). Córdoba, Argentina. Retrieved from:

http://www.investigacion.frc.utn.edu.ar/labsis/Publicaciones/congresos_labsis/cynthia/CNIT_2009_Aplicacion_Algoritmos_Weka.pdf

Demšar, J., Zupan, B., Leban, G., & Curk, T. (2004). Orange: From Experimental Machine Learning to Interactive Data Mining. In J.-F. Boulicaut, F. Esposito, F. Giannotti, & D. Pedreschi (Eds.), *Knowledge Discovery in Databases: PKDD 2004.* (Vol. 3202, pp. 537–539). Pisa, Italy: Springer, Berlin, Heidelberg. doi: https://doi.org/10.1007/978-3-540-30116-5_58

Gironés Roig, J., Casas Roma, J., Minguillón Alfonso, J., & Caihuelas Quiles, R. (2017). *Minería de datos: modelos y algoritmos*. Barcelona: Editorial UOC.

Goddard, J. C., Cornejo, J. M., Martínez, F. M., Martínez, A. E., Rufiner, H. L., & Acevedo, R. C. (1995). Redes Neuronales y Árboles de Decisión: Un enfoque híbrido. In *Simposium Internacional de Computación organizado por el Instituto Politécnico Nacional* (pp. 1–7). Retrieved from: http://sinc.unl.edu.ar/sinc-publications/1995/GMMCRA95/sinc_GMMCRA95.pdf

Gómez Fontanills, D. (2012). Panoràmica de la wikimediasfera. *Digithum: Las Humanidades En La Era Digital*, (14), 25–34. Retrieved from: https://www.raco.cat/index.php/Digit/article/view/254239/341166

Hand, D. J., Mannila, H., & Smith, P. (2001). *Principles of Data Mining*. Massachusetts: MIT Press.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. Retrieved from: https://www.cs.cmu.edu/~tom/pubs/Science-ML-2015.pdf

Keegan, B. C. (2013). A history of newswork on Wikipedia. In *Proceedings of the 9th International Symposium on Open Collaboration - WikiSym '13* (pp. 1–10). Hong Kong, China: ACM Press. doi: https://doi.org/10.1145/2491055.2491062

Layton, R., Watters, P., & Dazeley, R. (2012). Recentred local profiles for authorship attribution. *Natural Language Engineering*, *18*(3), 293–312. doi: https://doi.org/10.1017/S1351324911000180

Mehdi, M., Okoli, C., Mesgari, M., Nielsen, F. Å., & Lanamäki, A. (2017). Excavating the mother lode of human-generated text: A systematic review of research that uses the wikipedia corpus. *Information Processing & Management*, *53*(2), 505–529. doi: https://doi.org/10.1016/j.ipm.2016.07.003

Mitra, S., & Acharya, T. (2003). *Data mining : multimedia, soft computing, and bioinformatics*. New Jersey: John Wiley & Sons.

Molina, L. C. (2002). Data mining: torturando a los datos hasta que confiesen., 1–11. Retrieved from http://www.uoc.edu/molina1102/esp/art/molina1102/molina1102.html

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. Massachusetts: MIT press.

Naik, A., & Samant, L. (2016). Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime. *Procedia Computer Science*, *85*, 662–668. doi: https://doi.org/10.1016/J.PROCS.2016.05.251

Okoli, C., & Schabram, K. (2009). Protocol for a systematic literature review of research on the Wikipedia. In *Proceedings of the International Conference on Management of Emergent Digital EcoSystems - MEDES '09* (pp. 458–459). Lyon, France: ACM Press. doi: https://doi.org/10.1145/1643823.1643912

Overell, S., Sigurbjörnsson, B., & van Zwol, R. (2009). Classifying tags using open content resources. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining - WSDM '09* (pp. 64–73). Barcelona, Spain: ACM Press. doi: https://doi.org/10.1145/1498759.1498810

Pacuit, E., & Parikh, R. (2006). Social interaction, knowledge, and social software. In *Interactive Computation: The New Paradigm* (1st ed., pp. 441–461). Springer Berlin Heidelberg. doi: https://doi.org/10.1007/3-540-34874-3_17

Riquelme Santos, J. C., Ruiz, R., & Gilbert, K. (2006). Minería de datos: Conceptos y tendencias. *Inteligencia Artificial: Revista Iberoamericana de Inteligencia Artificial, 10 (29),* 11-18. Retrieved from: https://idus.us.es/xmlui/bitstream/handle/11441/43290/Miner%c3%ada%20de%20datos.pdf?sequence=1&isAllowed=y

Ristoski, P., Bizer, C., & Paulheim, H. (2015). Mining the Web of Linked Data with RapidMiner. *Journal of Web Semantics*, *35*, 142–151. doi: https://doi.org/10.1016/j.websem.2015.06.004

Sánchez-Jiménez, R. (2007). La documentación en el proceso de evaluación de Sistemas de Clasificación Automática. *Documentación de Las Ciencias de La Información*, *30*, 25–44. Retrieved from: https://dialnet.unirioja.es/servlet/articulo?codigo=2316530

Saorín, T. (2012). *Wikipedia de la A a la W.* Barcelona: Editorial UOC.

Saorín, T. (2017). Wikipedismo de actualidad. La enciclopedia escrita desde el periodismo. *Anuario ThinkEPI*, *11*(1), 191–199. doi: https://doi.org/10.3145/thinkepi.2017.35

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. doi: https://doi.org/10.1145/505282.505283

Sun, A., & Lim, E.-P. (2001). Hierarchical text classification and evaluation. In *Proceedings 2001 IEEE International Conference on Data Mining* (pp. 521–528). California, USA: IEEE Comput. Soc. doi: https://doi.org/10.1109/ICDM.2001.989560

Taraborelli, D. (2013). Descending mount everest. In *Proceedings of the 9th International Symposium on Open Collaboration - WikiSym '13*. Hong Kong, China: ACM Press. doi: https://doi.org/10.1145/2491055.2491093

Tramullas Saz, J. (1997). Perspectivas en recuperación y explotación de información electrónica: el "Data Mining." *Scire: Representación y Organización Del Conocimiento*, *3*(2), 73–84. Retrieved from: https://www.ibersid.eu/ojs/index.php/scire/article/view/1077

Tramullas, J. (2015). Wikipedia como objeto de investigación. *Anuario ThinkEPI*, *9*(1), 223–226. doi: https://doi.org/10.3145/thinkepi.2015.50

Wikimedia Foundation. (2019). Retrieved from: https://wikimediafoundation.org/es/?noredirect=es_ES

Zupan, B., & Demsar, J. (2008). Open-Source Tools for Data Mining. *Clinics in Laboratory Medicine*, *28*(1), 37–54. doi: https://doi.org/10.1016/j.cll.2007.10.002